FACULTAD DE CIENCIAS JURÍDICAS Y SOCIALES

APLICACIÓN DEL BIG DATA AL ANÁLISIS DE LA DELINCUENCIA: UN ESTUDIO DE CASO

Carlos Domenech María

Grado en Seguridad Pública y Privada

Dpto. Estadística, Matemáticas e informática

Tutora: Marina Leal Palazón

Elche, 2025

RESUMEN

Este Trabajo de Fin de Grado analiza la aplicación del Big Data al estudio de la criminalidad, a través de una aplicación con el conjunto de datos reales sobre crimen US Crime. El objetivo principal es determinar qué variables presentan una relación significativa con la tasa de delitos y construir un modelo de regresión lineal capaz de predecirla. Tras probar regresiones simples y múltiples con distintos subconjuntos de variables, el modelo que ofrece mejores resultados es una regresión lineal múltiple completa, que muestra una capacidad predictiva adecuada pese al error propio de fenómenos sociales complejos. El estudio demuestra el valor del análisis de datos y del Big Data para comprender patrones delictivos y apoyar la toma de decisiones en seguridad y defensa.

Palabras clave: Big Data, criminalidad, regresión lineal, análisis de datos, predicción.

ABSTRACT

This Final Degree Project analyzes the application of Big Data to the study of crime, through an application with the real crime data set US Crime. The main goal is to identify variables significantly related to crime rates and to build a linear regression model capable of predicting them. After testing simple and multiple regressions, the best-performing model is a full multiple linear regression, which shows reasonable predictive accuracy despite the inherent complexity of social phenomena. The study highlights the value of data analysis and Big Data as tools for understanding criminal patterns and supporting decision-making in security and defense.

Key Words: Big Data, crime, linear regression, data analysis, prediction.

TABLA DE CONTENIDOS

1. INTRODUCCIÓN 7

1.1.Contexto del tema. 7

1.2.Justificación y objetivo de la investigación. 10

1.3.Aportación personal y novedad del TFG. 11

2. MARCO TEÓRICO 13

2.1.El Big Data. 13

2.2.Aplicaciones del Big Data en seguridad y defensa. 19

2.3.Análisis de datos. 23

2.4.Aprendizaje automático. Supervisado vs no supervisado. 25

2.5.Regresion. Técnicas y métodos. 29

2.5.1 Regresión lineal simple. 30

2.5.2 Regresión lineal múltiple. 32

2.5.3 Hipótesis nula, hipótesis alternativa y p-valor. 33

3. ESTUDIO EMPÍRICO COMPUTACIONAL. APLICACIÓN A CASO REAL 36

3.1.Descripción del conjunto de datos. US Crime Dataset. 36

3.2.Aplicación de método de regresión en R. 38

3.3.Presentación de resultados regresión. 41

4. DISCUSION 68

4.1.Interpretación de los resultados. 68

4.2.Coherencia de los resultados con estudios previos 71

4.3.Implicaciones teóricas o prácticas. 71

4.4.Limitaciones del estudio. 72

5. CONCLUSIONES 73

5.1.Resumen de los resultados más relevantes. 73

5.2.Evaluación del cumplimiento de los objetivos. 74

5.3.Propuestas futuras o recomendaciones. 75

6. REFERENCIA BIBLIOGRÁFICA 76

LISTADO DE ABREVIATURAS

IA Inteligencia Artificial

ML Machine Learning

IoT Internet of Things (Internet de las cosas)

BD Big Data

RLS Regresión Lineal Simple

RLM Regresión Lineal Múltiple

1. INTRODUCCIÓN

1.1. Contexto del tema.

A lo largo de la historia, la humanidad ha experimentado una serie de transformaciones profundas que han marcado hitos fundamentales en su desarrollo económico, social y cultural. Cada etapa histórica ha estado impulsada por descubrimientos y avances tecnológicos que no solo han redefinido los sistemas productivos y las estructuras laborales, sino que también han modificado la organización social, los modelos económicos y la forma en que las personas se relacionan entre sí. Estas transiciones, conocidas como revoluciones industriales, han actuado como verdaderos catalizadores del progreso humano.

La Primera Revolución Industrial: la era de la mecanización

La Primera Revolución Industrial, se sitúa a mediados del siglo XVIII, marcó la primera etapa de la mecamnización entre lo que destacó el invento de la máquina de vapor por James Watt. Este avance tecnológico permitió reemplazar el trabajo realizado por las personas y por los animales por energía mecánica, impulsando la creación de fábricas y la producción en masa. La economía experimentó un cambio en su forma: se pasó de una etapa agraria a un sistema industrial. De eta forma se incrementó la productividad y se impulsó una notable expansión del comercio.

La Segunda Revolución Industrial: electrificación y producción en serie

La Segunda Revolución Industrial, se desarrolló entre finales del siglo XIX y comienzos del XX, su característica principal fue la incorporación de la electricidad, el motor de combustión interna y los primeros sistemas de producción en serie. Estos avances permitieron que todo fuera más eficiente en los procesos industriales, las tareas automatizadas y la fabricación masiva de bienes provocando que se redujera el coste. Posteriormente,apareciónel automóvil, el teléfono y el avión que produjeron un cambio en la movilidad, las

comunicaciones y la distribución de mercancías, contribuyendo a la consolidación de un mercado global.

La estandarización de productos y el la nueva forma de organizar el trabajo, optimizaron la productividad y establecieron las bases del capitalismo moderno.

La Tercera Revolución Industrial: la era digital y la información

La Tercera Revolución Industrial, iniciada en la segunda mitad del siglo XX, es reconocida como la Revolución Digital o la Era de la Información. Su rasgo distintivo fue la irrupción de la electrónica, la informática y, posteriormente, de internet, lo que permitió la digitalización progresiva de los procesos productivos y sociales.

La interconexión del mundo permitió el acceso inmediato a mucha más informacion, esto transformó de forma notable las formas de comunicación, la educación, el entretenimiento y el trabajo. El conocimiento se convirtió en el nuevo recurso estratégico y la información pasó a ser una parte principal de la economía.

En esta etapa surgieron también los primeros sistemas de producción autónoma y las bases de la inteligencia artificial moderna.

La Cuarta Revolución Industrial: la convergencia tecnológica y el Big Data

En el siglo XXI hemos podido observar como se ha consolidado la Cuarta Revolución Industrial, que también se conoce como la Industria 4.0. Esta nueva fase se caracteriza por el auge de tecnologías digitales, físicas y biológicas, que tienen como impulsor a la inteligencia artificial, el Internet de las Cosas (IoT), la robótica, la computación en la nube y el análisis de grandes cantidades de datos.

Hoy en día, la fabricación de las empresas, las ciudades y los hogares están interconectados de manera permanente. Los dispositivos inteligentes, los

sensores, los relojes, teléfonos, etc. generan cantidad infinita de información que circula a través de redes digitales. Este fenómeno es lo que se conoce como Big Data, un concepto que describe el manejo de grandes volúmenes de datos, muy diversos y sobre todo dinámicos que requieren nuevas técnicas de almacenamiento, procesamiento y posteriormente de análisis.

La interconexión total ha permitido que los procesos sean mucho más productivos, nos ha permitido también optimizar comportamientos y tomar decisiones en tiempo real. Sin embargo, el gran desafío de esta epoca no está tan solo en la recopilación de esa información, sino en la capacidad de procesarla, interpretarla y transformarla en un conocimiento que nos sea de utilidad.

Aquí es donde intervente el aprendizaje automático (machine learning), disciplina que permite a los sistemas que vayan aprendiendo de los datos y así poder mejorar su rendimiento sin intervención directa de ninguna persona. Gracias a todos estos algoritmos, es posible detectar patrones, predecir tendencias y automatizar tareas complicadas, pudiendo ser más eficientes e innovando en sectores tan distintos como puéden ser la salud, , la educación o la economía.

En este sentido, la Cuarta Revolución Industrial no solo crea un nuevo entorno productivo, sino también cambia la estructura de la sociedad contemporánea, marcada por el mundo iterconectado, la automatización de ciertos procesos y la inteligencia de los sistemas. La gestión adecuada del Big Data y su utilización en las nuevas tecnologías constituyen,, la clave del gran desarrollo de este siglo.

1.2.Justificación y objetivo de la investigación.

El estudio de la criminalidad es un ámbito muy importante para las administraciones públicas, ya que para ellas, el poder entender qué factores pueden influir en las variaciones de los delitos, permite poder realizar intervenciones más eficaces y ajustadas a nuestra realidad. En este contexto, las herramientas de análisis de datos y los métodos propios del Big Data ofrecen nuevas posibilidades para que podamos examinar fenómenos, permitiendo integrar mucha más información y así poder encontrarrelaciones que no serían tan evidentes utilizando otros métodos tradicionales.

Este Trabajo de Fin de Grado tiene su origen en la curiosidad por la aplicación de estas técnicas al análisis de la crimen con un enfoque empírico y cuantitativo. Para ello se ha empleado el US Crime Dataset, una base de datos de referencia que es bastante común en el estudio del delito, que reúne en ella variables económicas, demográficas y policiales correspondientes a distintos estados de Estados Unidos. La utilización de regresiones realizdas en el software R ha permitido evaluar la relación lineal entre estas variables y la tasa de criminalidad, ofreciendo una aproximación bastante buena.

El trabajo, además, ha permitido mostrar cómo los métodos estadísticos utilizados y el análisis de esa gran cantidad de datos pueden complementar las teorías de la criminología, proporcionado una base sólida para que se puedan implantar unas políticas públicas más adecuadas.

Objetivo general

Analizar qué variables del US Crime Dataset tienen una relación significativa con la tasa de criminalidad (Crime), mediante la aplicación de técnicas de regresión lineal en R e implementar un modelo de regresión lineal capaz de predecir la tasa de criminalidad.

Objetivos específicos

- Contextualizar el concepto de Big Data y analizar su utilidad en la criminología, destacando su papel en el estudio, gestión y predicción de fenómenos delictivos.
- Examinar y describir las variables contenidas en el conjunto de datos US Crime, identificando su relevancia.
- Aplicar regresiones lineales simples para determinar la influencia individual de cada predictor sobre la tasa de delitos.
- Seleccionar las variables con mayor relevancia estadística para su inclusión en un modelo más complejo.
- Construir un modelo de regresión lineal múltiple que permita analizar el efecto conjunto de los factores más relevantes.
- Realizar una predicción para evaluar el modelo de regresión lineal múltiple elegido.
- Interpretar los resultados obtenidos.
- Evaluar el potencial del Big Data y de las técnicas estadísticas como herramientas útiles para el estudio y comprensión de fenómenos delictivos.

1.3. Aportación personal y novedad del TFG.

En el Grado en Seguridad Pública y Privada he estudiado el fenómeno delictivo desde perspectivas criminológicas y jurídicas, estudiando que el crimen actual factores sociales, económicos, etc. Sin embargo, la realización de este TFG me ha llevado un paso más allá y ha supuesto un reto para mí ya que aunque no he cursado la asignatura de Big Data, siempre me ha suscitado gran interés en cómo todos esos datos que se recopilan a diario, pueden ayudar en el ámbito de la seguridad y la defensa.

La principal novedad que me aporta este trabajo es haber sido capaz de utilizar los conocimientos que he adquierido sobre la criminalidad con métodos estadísticos que no había utilizado casi durante la carrera.

Me he enfrentado a una base de datos real, he podido aprender a seleccionar las variables relevantes para luego decidir qué modelos aplicar y

extraer conclusiones. Todo ello ha representado un desafío que me ha permitido desarrollar un grado más avanzado de conocimiento para poder utilizar en mi desarrollo profesional. Para ello, he utilizado el software estadístico RStudio, una herramienta utilizada en el análisis de datos y que he tenido que aprender a manejar para poder realizar este trabajo. Asimismo, he empleado RMarkdown, una extensión que no conocía y que permite poder introducir en un documento el código, los resultados, y las gráficas, haciendo más sencillo que el análisis sea más ordenado.

Además, este TFG me ha hecho entender cómo el Big Data puede ayudar en la labor policial y la gestión de la seguridad, aportando una visión más objetiva, pudiendo actuar con anticipación y pudiendo tomar decisiones de forma más adecuada.. Todo esto ha ampliado mi perspectiva sobre los usos que pueden proporcionar las nuevas tecnologías para la seguridad pública y privada. En este sentido, el TFG ademas de haberme permitido utilizar lo que he aprendido durante la cerrara, también me ha permitido ver los futuros usos del Big Data, como la predicción de patrones delictivos, la optimización de recursos policiales o el diseño de estrategias de prevención basadas en datos reales.

En conjunto, este trabajo ha sido una parte más de mi formación y, al mismo tiempo, me ha hecho descubrir nuevas herramientas que pueden resultarme útiles en mi futuro profesional dentro del ámbito de la seguridad..

2. MARCO TEÓRICO

2.1. El Big Data.

Big Data es un término que se refiere a un gran conjunto de datos, generado tan rápidamente, que las herramientas tradicionales de almacenamiento, gestión y análisis no pueden manejarlo. No se trata solo de la que sean muchos datos los que se recopilan sino también de su diversidad, y de la gran velocidad con la que se capturan y posteriormente hay que hacer que sean útiles. Este concepto presenta un desafío que no había sido visto hasta la fecha y que requiere nuevas infraestructuras, arquitecturas distribuidas, algoritmos avanzados que puedan dotar de valor a los datos y a su vez otros y enfoques que nos permitan extraer valor de la información.

Diferentes instituciones y expertos han ofrecido definiciones complementarias para entender mejor este fenómeno. O'Reilly Radar menciona que "Big Data es cuando el volumen de los datos se convierte en parte del problema a resolver" (Dumbill, 2012), destacando que la enorme cantidad de información puede dificultar el uso de métodos convencionales y exigir el desarrollo de técnicas más sofisticadas.

Por otro lado, **IBM Big Data Platform** describe el concepto como "*activos de información de alto volumen, alta velocidad y/o alta variedad que requieren nuevas formas de procesamiento para mejorar la toma de decisiones, descubrir información y optimizar procesos*" (Gartner, 2012). Esta definición es muy citada en el ámbito empresarial y académico y dio origen a las conocidas "tres V": volumen, velocidad y variedad, que caracterizan al Big Data. Más tarde, se añadieron dos dimensiones más: la veracidad, que se refiere a la calidad y fiabilidad de la información, y el valor, que se relaciona con la capacidad de transformar los datos en conocimiento útil convirtiéndose así en las "Cinco Vs".

Desde un punto de vista más formal y académico, **De Mauro, Greco y Grimaldi** definen el Big Data como "*un activo informativo caracterizado por un alto volumen, velocidad y variedad que requiere tecnologías específicas y métodos*

analíticos para su transformación en valor" (De Mauro, Greco & Grimaldi, 2016). Esta definición resalta que los datos por sí solos no son útiles si no se tratan adecuadamente y se transforman en información procesable y significativa.

A estas definiciones se suma la de **Viktor Mayer-Schönberger y Kenneth Cukier**, autores de un texto influyente sobre el tema. Según ellos, "*Big Data no se trata solo de manejar grandes volúmenes de información, sino de un nuevo modo de comprender y analizar el mundo a través del procesamiento masivo de datos para encontrar patrones y correlaciones que antes eran invisibles*" (Viktor Mayer-Schönberger y Kenneth Cukier, 2013). Esta visión amplía el alcance del concepto, entendiéndolo no solo como un desafío técnico, sino como un cambioen la forma de genera el conocimiento, un cambio en cómo, se toman decisiones y se diseñan políticas en toda la sociedad.

El campo del Big Data ha evolucionado notablemente en estos últimos tiempos. Al principio, el análisis de datos se centraba en técnicas estadísticas aplicadas a bases de datos muy pequeñas y que se encontraban ordenadas. Hoy en día, los datos se encuentra repartidos en la nube y se necesitan técnicas de procesamiento paralelo, inteligencia artificial, aprendizaje automático (machine learning) y análisis predictivo. Este progreso ha tenido un gran impacto en la sociedad actual, donde el volumen de datos generados diariamente por redes sociales, dispositivos IoT, sensores, sistemas financieros o plataformas de servicios supera las capacidades que hasta ahora se tenían para procesar eso datos,

La capacidad de gestionar, analizar e interpretar estos datos se ha convertido en un elemento que persiguen empresas, gobiernos y organizaciones para hacerse más efectivas. Gracias al Big Data, las organizaciones pueden anticipar comportamientos, identificar tendencias, optimizar sus recursos y diseñar estrategias que les permiten adaptarse más rápidamente a los cambios. Este aprovechamiento de los datos, ademas de proporcionar a las empresas una ventaja en el mercado, hace que se propicie un nuevo modelo económico que se basa en una toma de decisiones informada y en el conocimiento derivado del análisis de grades cantidades de información.

En este sentido, los macrodatos, termino que se utilizar en español para referirnos al Big Data (BD), hace que el foco no sea solo el estudio de cada dato de forma individual sino que intenta extraerse el valor del estudio de los datos en conjunto. Se buscan patrones y relaciones significativas entre variables que nos permitan poder explicar fenómenos complejos y así poder tomar decisiones más acertadas. Estos datos, que pueden presentarse de muchas formas como por ejemplo, texto, imágenes, señales en tiempo real, son procesados mediante herramientas avanzadas que nos permiten transformar toda esa gran cantidad de información en una forma de conocimiento que podemos luego utilizar. Este enfoque facilita la toma de mejores decisiones basadas en la información, lo que está transformando la forma en la que la sociedad y las empresas aplican sus estrategias.

El Big Data no debe entenderse como una tecnología en si misma, sino como un conjunto orientado a la gestión, el análisis y la extracción de valor de grandes cantidades de información. La importancia de todo esto, está en la capacidad de transformar la información en conocimiento útil que nos sirva para poder tomar decisiones estratégicas en diversos ámbitos, como en el empresaria, el social y el gubernamental.

IMAGEN 1
Generada por ChatGPT (modelo GPT-5) de OpenAI

En este contexto, y tal como hemos visto antes, para poder trabajar con eficacia con estas grandes cantidades de datos, se definieron las dimensiones del Big Data que a continuación vamos a presentar. (IMAGEN 1).

VOLUMEN

La primera y más llamativa dimensión del Big Data es el volumen, que como podemos imaginar, hace referencia a la cantidad total de datos que se generan, almacenan y luego se procesan. En la actualidad, el crecimiento exponencial del uso de dispositivos conectados, redes sociales, plataformas

digitales, sensores del Internet de las Cosas (IoT) y sistemas transaccionales ha provocado que se produzcan grandes cantidades de datos en un tiempo muy limitado.

Estos datos pueden venir de multitud da lugares: interacciones en redes sociales, registros que se producen al comprar en línea, historiales médicos digitalizados, datos de geolocalización, imágenes, vídeos, mediciones de sensores, etc.

La importancia del volumen reside en que cuanto más datos disponibles, mayor será la posibilidad de identificar patrones importantes, tendencias y correlaciones entre los. Con conjuntos de datos grandes, los modelos analíticos y predictivos serán muchos más precisos y eso nos permitirá obtener mejores conclusiones. Además, el gran volumen favorece la aplicación de técnicas avanzadas como el aprendizaje automático (machine learning) o la inteligencia artificial, que requieren grandes cantidades de información para entrenarse adecuadamente como veremos a lo largo del trabajo.

VELOCIDAD

La velocidad se refiere a la rapidez con la que los datos se generan, transmiten, capturan y procesan. En esta época, esta característica ha adquirido una importancia crucial, ya que los datos se generan a gran velocidad y desde multitud de fuentes en tiempo real o con una gran rapidez..

Unos ejemplos de esto pueden ser, las operaciones financieras, los sistemas de control industrial, los sensores en vehículos autónomos o las interacciones en redes sociales, todos ellos generando información de manera continua y a mucha velocidad.

La capacidad de procesar esta información lo más rápidamente posible es fundamental porque una respuesta rápida puede marcar la diferencia entre el éxito y el fracaso en la toma de decisiones. Un ejemplo lo podemos encontrar, en el ámbito empresarial, el análisis en tiempo real permite a las empresas hacer un ajuste de sus estrategias de marketing sin demora, permite así mismo detectar

fraudes en el momento en que ocurren o personalizar la experiencia del cliente de manera inmediata. La tendencia actual es avanzar hacia sistemas capaces de gestionar datos en en tiempo real, es decir, mientras se producen, reduciendo al mínimo el tiempo entre la captura de la información y su análisis.

VARIEDAD

La variedad es otro aspecto clave, que se refiere a la diversidad de tipos y formatos de datos que necesitamos gestionar. A diferencia de los sistemas tradicionales que trabajaban principalmente con datos estructurados (organizados en tablas y bases de datos relacionales), el Big Data incorpora una amplia gama de datos no estructurados (como textos, vídeos, audios o publicaciones en redes sociales) y semiestructurados (como correos electrónicos o registros XML).

Esta gran cantidad de datos dispares representa un reto a valorar, ya que requiere técnicas de almacenamiento y análisis flexibles capaces de manejar datos con diferentes estructuras, de diferentes orígenes y con distintos niveles de calidad. No obstante, esto también genera una gran oportunidad, porque permite obtener una visión más completa de la realidad al integrar información que proviene de fuentes diferentes. La capacidad de correlacionar datos heterogéneos es, uno de los aspectos que dota al Big Data su mayor poder analítico.

VALOR

Destacar también como otra dimensión del Big Data, el valor. Este es el elemento que justifica la inversión en infraestructuras, tecnologías y recursos humanos para la gestión de datos. De nada sirve recopilar enormes cantidades de información; lo realmente importante es la capacidad de transformarla en un algo útil que nos permita toma decisiones acertadas..

El valor se manifiesta cuando los datos analizados permiten, por ejemplo, mejorar procesos, aumentar la eficiencia, reducir costes, diseñar productos personalizados o poder predecir comportamientos futuros. Esta dimensión nos

muestra que el objetivo final del Big Data no es almacenar información porque si, el objetivo ultimo es convertir esa información en en conocimiento aplicable que aporte ventajas competitivas y beneficie a las empresas y a la sociedad en general.

VERACIDAD

Por último, tenemos la veracidad. Este términohace referencia a la calidad, precisión, consistencia y fiabilidad de los datos. Dado que las decisiones que se toman, se basan en el análisis de esta información, resulta importante que los datos sean representativos, hayan sido depurados eliminando los errores y hayan sido validados adecuadamente. La presencia de datos incompletos, duplicados, sesgados o no válidos puede llevarnos a cometer interpretaciones erróneas y decisiones no beneficiosas.

Para garantizar la veracidad, es necesario implementar procesos de limpieza, verificación y control de calidad que nos garanticen que los datos utilizados sean fiables y tenga la significacion adecuada. Esto implica, entre otras cosas, la validación de las fuentes de los datos, la eliminación de valores que no nos sirvan, los no representativos y el mantenimiento de protocolos estandarizados en la recolección y en el uso de la información.

En conjunto, las cinco “Vs” del Big Data constituyen el marco conceptual que permite entender su complejidad y a su vez su potencial transformador. Cada una de estas dimensiones es esencial para extraer el máximo valor de los datos en entornos tan cambiantes y complicados. El éxito de cualquier estrategia que se base en el Big Data va a depender, mayoritariamente, de la capacidad para gestionar adecuadamente estos cinco aspectos, que actúan de manera interdependiente.

Solo mediante la integración de volumen, velocidad, variedad, valor y veracidad es posible aprovechar todo el potencial que ofrece el Big Data y convertir la información en un recurso estratégico que impulse la innovación, la eficiencia y la toma de decisiones informadas.

Por todo ello, podemos concluir, que el término Big Data se refiere al manejo de volúmenes masivos de datos generados continuamente en diversos formatos, que exceden las capacidades de los sistemas tradicionales de almacenamiento y procesamiento (Volumen). No se trata tan solo de la cantidad de información, sino también de su velocidad de creación y su heterogeneidad (Velocidad y Variedad). Todo este proceso tan complejo, requiere infraestructuras tecnológicas avanzadas para almacenaje y métodos de análisis especializados para procesar, interpretar y extraer conocimiento de estos datos (Veracidad). Como bien hemos dicho ya antes, el objetivo del Big Data no es la acumulación de información, sino su transformación en valor a través del análisis (Valor). Esto permite mejorar la toma de decisiones, anticiparse y realizar predicciones adecuadas, optimizar procesos y descubrir patrones que sería imposible identificar con técnicas convencionales. En resumen, Big Data se presenta como un reto técnico y provoca un cambio en la forma de entender el mundo que hasta ahora conocemos generando conocimiento basado en evidencias que se derivan de los grandes volúmenes de datos.

2.2. Aplicaciones del Big Data en seguridad y defensa.

El desarrollo de las tecnologías y en particular, des Big Data, ha revolucionado la forma en que los Estados y las instituciones toman en cuenta temas tan importantes como la seguridad y la defensa. Su capacidad para procesar, integrar y analizar vastas cantidades de información en tiempo real permite anticipar amenazas, coordinar respuestas más eficientes y optimizar el uso de los recursos. El Big Data se ha configurado en cuanto la gestión de riesgos, la prevención del crimen y la toma de decisiones estratégicas, con una herramienta fundamental.

Seguridad pública y análisis predictivo

En el ámbito de la seguridad ciudadana, el Big Data se emplea para desarrollar sistemas que analizan grandes cantidades de información con el objetivo de identificar patrones de comportamiento y tendencias o

comportamientos delictivos. A través de algoritmos, todos los datos que se van recopilando mediante cámaras de videovigilancia, sensores urbanos, redes sociales o bases de datos policiales son procesados para detectar zonas o situaciones en las que se pudiera dar una mayor probabilidad de riesgo.

Esta capacidad de hacer predicciones permite a las fuerzas y cuerpos de seguridad anticiparse a posibles delitos, orientar sus actuaciones y optimizar la distribución de los recursos de seguridad. Actuando de manera preventiva, las administraciones pueden aumentar la eficacia de sus intervenciones siendo de esta forma más eficientes. No obstante, esto también conlleva numerosos desafíos sobre todo en lo que respecta a la privacidad y al uso de datos personales.

IMAGEN 2. NYPD and Microsoft collaborate to create the Domain Awareness System (DAS). Fuente: https://medium.com/homeland-security/nypd-and-microsoft-collaborate-to-create-the-domain-awareness-system-das-543a6245cb8f

Defensa e inteligencia estratégica

En el ámbito de la defensa, el Big Data es básicamente imprescindible, sobre todo para la obtención de datos para inteligencia y la planificación de operaciones militares. Al integrar datos de satélites, sensores, radares y sistemas de comunicación, se obtiene una visión más precisa del entorno operativo permitiendo tener cierta ventaja. Este análisis permite la detección temprana de amenazas, la evaluación de riesgos y la optimización de los recursos.

La aplicación de estas tecnologías también ha impulsado la creación de centros de mando y puestos de control avanzados capaces de unificar información recopilada en diferentes lugares. Gracias a ellos, los responsables de defensa pueden supervisar y coordinar operaciones en tiempo real, mejorando la capacidad de reacción ante situaciones críticas. No debemos olvidar tampoco que , el Big Data permite procesar información no estructurada como textos, imágenes o señales y transformarla en conocimiento que nos resulte útil. De esta forma, se refuerza la capacidad de análisis y da cierta ventaja en la toma de decisiones.

Ciberseguridad y protección de infraestructuras críticas

La digitalización de los sistemas de seguridad y defensa ha convertido la ciberseguridad en un problema de primer orden. Las herramientas de Big Data permiten analizar a la vez, millones de registros de red para detectar anomalías, intrusiones o comportamientos sospechosos.

En el ámbito de la defensa nacional, estas tecnologías se aplican para proteger infraestructuras críticas, garantizar la seguridad de la información y prevenir ataques informáticos. Su habilidad para conectar datos de distintas fuentes nos ayuda a detectar problemas potenciales antes de que se conviertan en vulnerabilidades y a poner en marcha soluciones automáticas de forma rápida y eficiente. Este enfoque fortalece la seguridad digital y también contribuye a la estabilidad del conjunto del sistema de defensa.

De la reacción a la anticipación

El uso del Big Data en la seguridad y la defensa representa un cambio en la forma de hacer las cosas. Las estrategias tradicionales que se centraban en reaccionar ante los hechos una vez se hubieran producido producidos, han evolucionado hacia modelos basados principalmente en la anticipación y la prevención.

La posibilidad de analizar información en tiempo real, ver patrones que antes resultaban imperceptibles y poder generar predicciones permite tomar decisiones basadas en a la información antes de que se materialicen las amenazas. Esto se traduce en una gestión mucho más eficiente, una reducción de los riesgos y una mejora en la protección de los ciudadanos y del territorio.

Sin embargo, el aprovechamiento de estas ventajas debe acompañarse de una gestión responsable de la información, teniendo en cuenta que para que funcionen adecuadamente se debe garantizar la privacidad, la transparencia y el respeto a los derechos inherentes de las personas.

Aplicación real del Big Data en el ámbito de seguridad y emergencias

En el campo de la seguridad y la gestión de emergencias, uno de los ejemplos a tener en cuenta en cuanto a la aplicación del Big Data lo podemos encontrar en la ciudad de Nueva York. Allí se ha desarrollado un Centro de Excelencia (Domain Awareness System) desarrollado conjuntamente por el Departamento de Policía de Nueva York (NYPD) y Microsoft (IMAGEN 2), que se decida a la integración y análisis de información procedente de una gran variedad de fuentes. Este centro permite recopilar datos de cámaras, sensores urbanos, sistemas de tráfico y multitud de registros para posteriormente procesarlos y compartirlos de forma automática y en tiempo real con otros organismos públicos.

El objetivo principal de este centro es mejorar la coordinación entre. Instituciones y facilitar la toma de decisiones en tiempo real ante situaciones de riesgo, emergencias o amenazas a la seguridad ciudadana. Al tener disponibilidad de todos los datos, las autoridades pueden actuar con mayor rapidez y eficacia, optimizando los recursos disponibles y siendo así mucho más eficientes en la respuesta de los diferentes escenarios que pudieran surgir.

Además, el uso de herramientas de análisis de grandes cantidades de datos, nos permite detectar patrones y tendencias en los datos, lo que contribuye a anticipar incidentes y planificar estrategias preventivas. Este enfoque demuestra cómo el Big Data puede convertirse en un instrumento de gestión inteligente, capaz de transformar la gran cantidad de información que se va recopilando en conocimiento útil para hacer más eficaz las actuaciones públicas. El centro gestiona más de nueve mil cámaras de seguridad, sensores de radiación, sistemas de reconocimiento de matrículas, bases de datos y redes de transporte. Además, cuenta con una plataforma de análisis visual y predictivo en tiempo real.

Como podemos observar, la experiencia del Centro de Excelencia de Nueva York refleja el potencial que puede tener el Big Data para modernizar los sistemas de seguridad urbana, haciendo de la administración un servicios más proactiva, eficiente y basada en evidencias extraídas de los datos.

2.3.Análisis de datos.

El análisis de datos está formado por las técnicas y procesos utilizados para examinar, depurar, transformar e interpretar datos, con el objetivo de descubrir información significativa, extraer conclusiones y apoyar la toma de decisiones. Hasta ahora, este análisis se hacía sobre volúmenes relativamente pequeños de datos estructurados utilizando herramientas estadísticas clásicas. El crecimiento exponencial de la información, impulsado por internet, redes sociales, sensores, dispositivos móviles e inteligencia artificial, ha ampliado tanto su alcance como la dificultad de análisis.

En el mundo del Big Data, el análisis de datos es fundamental. No tiene sentido generar una gran cantidad de datos porque si. Lo que realmente los convierte en un recurso valioso es el poder analizarlos y extraer un conocimiento que nos sea de utilidad. Digamos que el análisis de datos es como una lupa que nos ayuda a reordenar el rompecabezas de los datos. Si no se procede al análisis de los datos, no deja de ser un mero conjunto de datos. Con su análisis se transforma en una nueva herramienta de apoyo a la toma de decisiones. El análisis de datos dentro del Big Data permite entre otras cosas:

- Descubrir patrones y tendencias ocultos entre grandes volúmenes de información.
- Predecir comportamientos futuros mediante modelos estadísticos o de aprendizaje automático.
- Personalizar servicios y mejorar la experiencia del usuario en función de sus datos y hábitos.

En palabras de Provost y Fawcett (2013), “el valor de los datos no reside en su existencia, sino en la capacidad de analizarlos para generar conocimiento”, lo que demuestra que el análisis es el puente que transforma información en utilidad.

ETAPAS DEL ANÁLISIS DE DATOS

Como hemos explicado antes, depuesto de recopilar los datos provenientes de distintas fuentes como sensores, redes sociales, transacciones y bases de datos, etc., se inicia el proceso de análisis de datos, el cual se estructura en las siguientes fases:

1. Limpieza y preparación: Depuración de errores, tratamiento de valores, eliminación de duplicados y transformación de los datos en formatos que se puedan analizar.
2. Exploración y visualización: Análisis preliminar para comprender la estructura de los datos, sus distribuciones y posibles patrones o datos erróneos..
3. Modelado y análisis estadístico: Aplicación de técnicas estadísticas, minería de datos o algoritmos de aprendizaje automático para identificar relaciones entre variables o poder predecir resultados.
4. Interpretación de resultados: Extracción de conclusiones y validación de las hipótesis planteadas.

Después de la finalización de las fases anteriormente mencionadas, se procede a la comunicación y posteriormente se procede a la toma de decisiones. Por último, se presentan los hallazgos a través de informes para facilitar su comprensión y aplicación.

El análisis de datos ha evolucionado de ser una simple herramienta de apoyo a convertirse en un componente estratégico clave para organizaciones de todo tipo, tanto públicas como privadas. Las empresas lo utilizan para optimizar sus cadenas de suministro, predecir la demanda de sus productos, detectar fraudes y diseñar campañas de marketing personalizadas que lleguen al público adecuado. Los gobiernos, por su parte, lo aplican para planificar políticas públicas de manera más eficiente, gestionar sus recursos de forma óptima y mejorar los servicios que ofrecen a los ciudadanos. En resumen, el análisis de datos se ha convertido en una herramienta poderosa que impulsa el progreso en muchos sectores.

La evolución del análisis de datos, ha ido desde métodos descriptivos hasta técnicas más complejas como el aprendizaje automático (machine learning), donde los algoritmos aprenden automáticamente a partir de los datos para hacer predicciones o clasificaciones.

2.4. Aprendizaje automático. Supervisado vs no supervisado.

Como hemos visto anteriormente, el Big Data hace referencia al manejo y análisis de gran cantidad de información que superan las capacidades de los sistemas tradicionales para ser procesados de manera eficiente. No se trata solo de la cantidad de datos, sino también de su volumen, velocidad, variedad, valor y veracidad, las conocidas “5 Vs” del Big Data. Por todo ello, el aprendizaje automático se ha convertido en una herramienta indispensable para extraer conocimiento útil de toda la información recopilada.

Los algoritmos de machine learning (ML), son los encargados de analizar automáticamente la gran cantidad de información para detectar patrones, realizar predicciones o generar modelos explicativos. Sin embargo, el tipo de aprendizaje empleado, supervisado o no supervisado, dependerá del tipo de datos disponibles y de los objetivos del análisis.

El aprendizaje automático tiene como objetivo dotar a las máquinas de la capacidad de aprender patrones o relaciones a partir de datos, sin necesidad de programar cada regla. Es decir, en lugar de codificar manualmente todas las decisiones, se le proporciona al sistema un conjunto de ejemplos (datos), y el algoritmo ajusta sus parámetros para generalizar ese conocimiento a nuevos casos (IBM, 2023). Dentro de ese amplio campo existen dos modalidades principales:

APRENDIZAJE SUPERVISADO

En entornos de Big Data, el aprendizaje supervisado se utiliza principalmente para predicciones precisas y clasificaciones automáticas a gran escala. Sus características principales son:

- En el aprendizaje supervisado, los datos con los que entrena el sistema ya contienen respuestas correctas (etiquetas o valores reales), de modo que el algoritmo “aprende” a asociar las entradas con las salidas esperadas.
- Durante el entrenamiento, el modelo realiza predicciones sobre los datos de entrenamiento, las compara con las respuestas verdaderas, calcula el error, y ajusta sus parámetros para minimizar ese error.
- Las tareas típicas incluyen clasificación (predecir una categoría discreta) y regresión (predecir un valor continuo).

Por ejemplo, si tienes un conjunto de correos electrónicos ya catalogados como spam o no spam, puedes entrenar un clasificador supervisado para que, pueda etiquetar correctamente los correos nuevos.

La principal ventaja del aprendizaje supervisado en el ámbito del Big Data es la gran capacidad de generalización, siempre que se cuente con conjuntos de datos amplios y correctamente etiquetados. Cuantos más ejemplos y casos prácticos de entrenamiento reciba el algoritmo, más precisa y fiable será la construcción del modelo para realizar futuras predicciones. Esto nos permitirá realizar estimaciones con un margen de error cada vez más pequeño. Esta característica es especialmente útil en aplicaciones como la detección de fraudes, el reconocimiento de imágenes o la predicción de comportamientos de los usuarios, donde la precisión es una parte esencial para el éxito del sistema.

A pesar de estas ventajas, uno de los principales desafíos sigue siendo la obtención de etiquetas de calidad a gran escala. El proceso de etiquetar manualmente enormes volúmenes de datos es complicado, consume mucho tiempo y requiere recursos humanos, lo que puede incrementar los costes de desarrollo de un proyecto de inteligencia artificial. Aunque existen técnicas como la semi-supervisión, el aprendizaje activo o el uso de herramientas de etiquetado automatizado que buscan minimizar este esfuerzo, todavía resulta complicado alcanzar la misma fiabilidad que ofrecen las etiquetas generadas por expertos.

Por lo tanto, encontrar un equilibrio entre la disponibilidad de datos correctamente etiquetados y la eficiencia en la creación de modelos predictivos

sigue siendo un objetivo principal para el desarrollo de soluciones basadas en Big Data y aprendizaje supervisado.

APRENDIZAJE NO SUPERVISADO

El aprendizaje no supervisado posee una gran relevancia en Big Data porque permite analizar información sin necesidad de haber unas etiquetas previas. En un entorno donde los datos crecen de forma exponencial y desestructurada esta característica resulta esencial.

Gracias a técnicas como el clustering o la reducción de dimensionalidad, los algoritmos no supervisados pueden descubrir patrones de comportamiento entre millones de usuarios sin intervención humana, detectar anomalías en tiempo real dentro de flujos de datos, segmentar mercados o grupos de clientes en función de sus características comunes, etc.

El aprendizaje no supervisado actúa como una herramienta de exploración y descubrimiento dentro del ecosistema del Big Data, ayudando a las organizaciones a encontrar relaciones y tendencias que serían imposibles de identificar con métodos manuales. Sus principales características son:

- El aprendizaje no supervisado trabaja con datos sin etiquetas: el modelo no recibe respuestas previas. Su misión es descubrir estructuras, agrupamientos o regularidades intrínsecas del conjunto de datos.
- No hay un “error” frente a una etiqueta correcta, sino funciones de optimización internas que guían al modelo a encontrar distribuciones, densidades o representaciones latentes del dato.
- Algunas de sus tareas frecuentes son clustering (agrupar objetos similares), reducción de dimensión (simplificar la representación de los datos), o detección de anomalías.

Por ejemplo, un sistema podría analizar los patrones de consumo de los clientes sin que nadie los haya clasificado previamente, y agrupar a los usuarios en segmentos con comportamientos similares.

A continuación, se presenta una comparación de las fortalezas y debilidades del aprendizaje supervisado frente al no supervisado:

Aspecto	Ventajas del aprendizaje supervisado	Ventajas del aprendizaje no supervisado
Precisión en predicciones / control	Al contar con etiquetas, puede optimizar directamente para el objetivo deseado, obteniendo resultados más ajustados al problema (mayor precisión).	No depende de datos etiquetados, por lo que puede aprovechar grandes volúmenes de datos sin tratar sin necesidad de costear su etiquetado.
Interpretabilidad / objetivo claro	El modelo tiene un objetivo definido y evaluable (por ejemplo, tasa de error, precisión, etc.), lo que facilita medir el desempeño y comparar.	Permite descubrir patrones no previstos o relaciones ocultas que no estaban definidas a priori por el investigador.
Requerimientos de datos	Si ya dispones de datos bien etiquetados, el enfoque supervisado es muy eficiente y directo.	Apoya en escenarios donde etiquetar es costoso, difícil o imposible, pues no exige etiquetas previas.
Aplicabilidad	Muy útil cuando el problema tiene una salida conocida (por ejemplo, diagnóstico médico, predicción financiera, clasificación de imágenes).	Excelente para exploración, segmentación, descubrimiento de estructuras nuevas o reducción de dimensiones.

Limitaciones del aprendizaje supervisado:

- Requiere disponer de una gran cantidad de datos etiquetados, lo cual puede ser costoso, lento o sujeto a errores humanos.
- Puede caer en sobreajuste: aunque el modelo encaje muy bien los datos de entrenamiento, su generalización a datos nuevos puede ser deficiente.
- Falta de flexibilidad para descubrir patrones nuevos fuera de las etiquetas estipuladas: el conocimiento que puede adquirir está limitado al dominio explícitamente definido.

Limitaciones del aprendizaje no supervisado:

- Al no existir etiquetas, es más difícil evaluar formalmente cuán “bueno” es el resultado; las métricas tradicionales no siempre aplican.
- Los resultados pueden no tener sentido práctico si los patrones detectados no son interpretables o relevantes para el contexto.
- Sensibilidad al ruido, valores atípicos o malas escalas de las características: estos pueden sesgar la agrupación o la estructura identificada.
- En algunos casos, los algoritmos no supervisados pueden generar agrupamientos triviales o poco útiles.

2.5.Regresion. Técnicas y métodos.

La regresión es una técnica del aprendizaje supervisado que se utiliza para poder predecir un valor numérico a partir de una o más variables de entrada.

“La regresión es la rama de la estadística en la que una variable dependiente de interés se modela como una combinación de una o más variables predictoras, junto con un término de error aleatorio.”

El objetivo final es poder encontrar una función que describa la relación entre variables independientes (X) y la variable dependiente (Y) que es la que se quiere predecir.

Un ejemplo podría ser el siguientes: si se dispone de datos sobre el número de patrullas, el tiempo de respuesta ante un aviso y el nivel de iluminación en zonas urbanas, junto con el número de delitos registrados, la regresión puede utilizarse para estimar cuales son los factores que influyen más en la reducción de hechos delictivos Esto permitiría poder optimizar la asignación de recursos policiales y mejorar las estrategias de prevención para incrementar la seguridad ciudadana.

Aunque existen muchos métodos de regresión, en el objeto de estudio de este TFG, vamos a abordar la regresión lineal y dentro de ella, sus versiones simple y múltiple.

La regresión lineal constituye uno de los métodos más sencillos y fundamentales dentro del aprendizaje supervisado. Se utiliza principalmente para predecir una variable cuantitativa a partir de una o más variables por medio de una relación lineal que es el rasgo que lo diferencia del del resto de regresiones. A pesar de su simplicidad en comparación con técnicas estadísticas más modernas, sigue siendo una herramienta de gran valor y aplicación frecuente en numerosos campos.

El modelo de regresión lineal se fundamenta en la idea de ajustar una relación lineal entre una variable respuesta y una o varias variables predictoras. Este enfoque permite interpretar relaciones entre variables y realizar predicciones cuantitativas de manera sencilla y eficiente.

2.5.1 Regresión lineal simple.

La regresión lineal simple es una técnica estadística utilizada para analizar y modelar la relación entre dos variables cuantitativas: una variable independiente (X, variable explicativa) y una variable dependiente (Y, variable respuesta) mediante una relación lineal. Su finalidad es determinar hasta qué punto los cambios en la variable independiente se asocian con variaciones en la dependiente, y permitir la predicción de valores futuros o desconocidos de esta última.

El modelo asume que la relación entre ambas variables puede representarse mediante una línea recta, cuya ecuación general es:

$$Y=a+bX+\varepsilon$$

En este modelo, Y representa la variable dependiente, es decir, el valor que se pretende estimar o predecir. X corresponde a la variable independiente, aquella

que influye o explica el comportamiento de Y. El parámetro a, conocido como ordenada en el origen, indica el valor de Y cuando X es igual a 0. El parámetro b define la pendiente de la recta y muestra cuánto varía Y por cada unidad de cambio en X. Finalmente, ε (épsilon) representa el término de error, que recoge las diferencias entre los valores observados y los estimados por el modelo, incluyendo los factores no explicados o no contemplados en la regresión.

La regresión lineal simple busca una línea que se ajuste lo mejor posible a una nube de puntos. Uno de los criterios utilizados para encontrar la mejor recta es la regla de mínimos cuadrados, que minimiza la distancia entre los valores observados (los puntos) y los valores ajustados (la recta).

COEFICIENTE DE CORRELACIÓN LINEAL

El coeficiente de correlación lineal, normalmente representado por la letra r, sirve para medir cómo se relacionan linealmente dos variables entre sí. En otras palabras, indica si cuando una variable cambia, la otra tiende a hacerlo también y en qué dirección. El valor de r siempre está entre –1 y +1.

- Si r está cerca de +1, significa que existe una relación lineal positiva fuerte: cuando una variable aumenta, la otra también lo hace.
- Si r está cerca de –1, la relación lineal es negativa: al aumentar una, la otra tiende a disminuir.
- Si r está cerca de 0, quiere decir que no hay una relación lineal clara entre ambas variables.

En la regresión lineal simple, este coeficiente ayuda a saber hasta qué punto la variable independiente (la que usamos para predecir) está realmente relacionada con la variable dependiente (la que queremos explicar linealmente). Además, si elevamos *r* al cuadrado (r^2), obtenemos una medida que nos dice qué parte de la variación de la variable dependiente puede explicarse por la variable independiente.

COEFICIENTE DE DETERMINACIÓN

El coeficiente de determinación, o r^2, es una medida que indica la eficacia con la que una recta de regresión se ajusta a los datos observados. En esencia, muestra qué proporción de la variabilidad en la variable dependiente puede explicarse por la variabilidad en la variable independiente.

Su valor siempre se encuentra entre 0 y 1. Cuando r^2 está cerca de 1, significa que el modelo logra representar de manera bastante precisa el comportamiento de los datos. En cambio, si su valor se aproxima a 0, el modelo apenas explica la relación entre las variables.

En el caso de la regresión lineal simple, r^2 se calcula elevando al cuadrado el coeficiente de correlación (*r*). De este modo, ofrece una idea clara de la calidad del ajuste del modelo y de la fuerza de la relación lineal entre las dos variables analizadas.

En síntesis, el coeficiente de correlación es más útil cuando se busca conocer si existe una relación lineal entre dos variables y en qué sentido ocurre, es decir, si la relación es positiva o negativa. En cambio, el coeficiente de determinación resulta más apropiado cuando el interés se centra en evaluar la capacidad del modelo para explicar la variabilidad de los datos.

En los análisis de regresión, ambos coeficientes aportan información complementaria, aunque el coeficiente de determinación (r^2) suele tener mayor relevancia, ya que permite valorar de forma más directa la calidad del ajuste del modelo.

2.5.2 Regresión lineal múltiple.

La regresión lineal múltiple es una extensión de la regresión lineal simple que permite analizar la relación entre una variable dependiente y varias variables independientes al mismo tiempo. Su objetivo es determinar cómo influye cada una de esas variables en el comportamiento de la variable que se quiere explicar.

El modelo se expresa mediante una ecuación del tipo:

$$Y=a+b_1X_1+b_2X_2 +\cdots+b_nX_n +\varepsilon$$

donde Y es la variable dependiente, “a” es la ordenada en el origen, $b_1,b_2,...,b_n$ son los coeficientes que indican el peso o influencia de cada variable independiente ($X_1,X_2,\ldots,X_n$), y ε representa el error o la parte de la variación de Y que el modelo no logra explicar.

Este tipo de regresión permite predecir un resultado o poder entender un fenómeno considerando varios factores a la vez, ofreciendo una visión más completa que la regresión simple, donde solo se analiza una variable independiente.

En la regresión lineal múltiple se utiliza principalmente el coeficiente de determinación (R^2), ya que permite valorar la capacidad explicativa global del modelo.

2.5.3 Hipótesis nula, hipótesis alternativa y p-valor.

El contraste de hipótesis es uno de los procedimientos que más se utilizan en el análisis estadístico. Nos permite determinar si los datos aportan evidencia suficiente para apoyar o rechazar una hipótesis. Este proceso se articula en torno a dos proposiciones opuestas: la hipótesis nula (H_0) y la hipótesis alternativa (H_1) (Casella y Berger, 2002).

La hipótesis nula (H_0) representa la situación de “no efecto”. Bajo esta hipótesis se asume que cualquier variación que podamos observar en los datos se debe al azar y no a un fenómeno real. H_0 expresa que no existe relación entre variables o que un tratamiento no produce cambios significativos. Por otra parte, podemos encontrar la hipótesis alternativa (H_1). Ésta recoge la posibilidad opuesta: que sí existe un efecto, una diferencia o una relación estadísticamente significativa.

El p-valor es la herramienta principal para decidir entre H_0 y H_1. Se define como la probabilidad de obtener un resultado igual o más extremo que el observado, bajo el supuesto de que H_0 es cierta. Un p-valor pequeño indica que los datos son poco compatibles con H_0, lo que proporciona evidencia a favor de H_1 (McLeod, 2025).

En el análisis estadístico, los niveles de significación 0,05, 0,01 y 0,001 se utilizan como umbrales para determinar si los resultados proporcionan evidencia suficiente para rechazar la hipótesis nula. Estos valores ayudan a clasificar la fuerza de la evidencia estadística. Un p-valor inferior a 0,05 se considera significativo, por debajo de 0,01 muy significativo y, si es menor que 0,001, altamente significativo.

APLICACIÓN DEL CONTRASTE DE HIPÓTESIS EN LA REGRESIÓN LINEAL

En la regresión lineal, el contraste de hipótesis se utiliza para comprobar si las variables que incluimos en el modelo realmente influyen en la variable que queremos explicar. Es decir, nos ayuda a saber si una variable tiene un efecto real o si el resultado puede deberse simplemente al azar. Para cada variable independiente se plantea lo siguiente:

- Hipótesis nula (H_0): la variable no tiene efecto. Esto significa que, aunque la incluyamos en el modelo, no aporta información ni ayuda a predecir la variable dependiente.
- Hipótesis alternativa (H_1): la variable sí tiene efecto. En este caso, la variable contribuye de forma significativa a explicar el comportamiento de la variable dependiente.

Para decidir entre estas dos opciones se utiliza el p-valor. Este valor nos dice si el efecto observado en los datos es suficientemente grande como para pensar que no es fruto del azar.

- Si el p-valor es pequeño (por debajo del nivel de significación, como 0,05), se considera que la variable sí influyen el modelo.

- Si el p-valor es grande, no podemos asegurar que la variable tenga un efecto real.

En resumen, el contraste de hipótesis en la regresión lineal sirve para saber qué variables son realmente importantes y si el modelo, en general, ofrece

3. ESTUDIO EMPÍRICO COMPUTACIONAL. APLICACIÓN A CASO REAL

3.1. Descripción del conjunto de datos. US Crime Dataset.

Para el estudio computacional que vamos a realizar en este TFG, vamos a usar el conjunto de datos US Crime Dataset. Se trata de una colección de datos que se encuentran organizados en forma de tabla con filas y columnas. Cada fila representa un estado de los Estados Unidos de América y cada columna representan los valores de las observaciones para cada estado.

La base de datos se construyó con motivo del interés de los criminólogos en encontrar la relación entre las penas impuestas y la tasa de criminalidad. En la década de los sesenta, Isaac Ehrlich recopiló los datos de los 47 estados de EE.UU utilizando las estadísticas del FBI y otros órganos gubernamentales. Su función principal era deducir la tasa de criminalidad. Para ello, analizó distintos factores sociales, económicos y penales intentando buscar relación con la tasa de criminalidad.

En 1973, publico su artículo “Participation in ilegitimate activities: A theroretical and empirical investigation en la revista Journal of Political Economy. En el artículo, basándose en las ideas anteriores de Gary Becker (1968), Ehrlich pudo verificar empíricamente la idea de que cometer delitos puede analizarse como una decisión racional, en la que las personas sopesan costos y beneficios antes de actuar (Teoría económica del delito). Intentaban relacionar las políticas de castigo como método disuasorio sobre la delincuencia. El conjunto de datos fue posteriormente depurado:

- **Ehrlich, I. (1973)** – Participation in illegitimate activities: a theoretical and empirical investigation, Journal of Political Economy, 81, 521–565.

- **Vandaele, W. (1978)** – Participation in illegitimate activities: Ehrlich revisited, en Deterrence and Incapacitation, Academia Nacional de Ciencias, EE.UU.

- **Venables, W. y Ripley, B. (1998)** – Modern Applied Statistics with S-Plus (2ª edición), Springer-Verlag.

El conjunto de datos empleado hoy en día constituye una versión redondeada y adaptada de los datos publicados por Vandaele (1978).

En el conjunto de datos podemos observar 16 variables, demográficas, económicas y penales, que las resumimos en la siguiente tabla:

Variable	Descripción	Tipo
M	Porcentaje de hombres de 14 a 24 años en la población total del estado	Númerica
So	Variable indicadora para un estado del sur	Categórica Binaria
Ed	Años promedio de escolarización de la población de 25 años o más	Numérica
Po1	Gasto per cápita en protección policial en 1960	Numérica
Po2	Gasto per cápita en protección policial en 1959	Numérica
LF	Tasa de participación laboral de hombres civiles urbanos de 14 a 24 años	Numérica
M.F	Número de hombres por cada 100 mujeres	Numérica
Pop	Población del estado en 1960 (en cientos de miles)	Numérica
NW	Porcentaje de personas no blancas en la población	Numérica
U1	Tasa de desempleo de hombres urbanos de 14 a 24 años	Numérica
U2	Tasa de desempleo de hombres urbanos de 35 a 39 años	Numérica
wealth	Valor mediano de los activos transferibles o del ingreso familiar	Numérica
Ineq	Desigualdad de ingresos: porcentaje de familias que ganan menos de la mitad del ingreso mediano	Numérica
Prob	Probabilidad de encarcelamiento: relación entre el número de condenas y el número de delitos	Numérica
Time	Tiempo promedio (en meses) que los delincuentes cumplen en prisión estatal antes de su primera liberación	Numérica
Crime	Tasa de criminalidad: número de delitos por cada 100,000 habitantes en 1960	Numérica

3.2.Aplicación de método de regresión en R.

La criminalidad, es uno de los aspectos sociales que más han sido estudiados. Para ello se utilizan diferentes perspectivas como pueden ser la sociológica, la psicológica, la jurídica y la económica. Cómo hemos visto antes, la teoría económica del delito que desarrollo Gary Becker y que posteriormente amplió Isaac Ehrlich, nos explicaba que las personas tomaban la decisión de delinquir valorando los beneficios frente al costo o castigó que ello podía acarrearles.

En la línea que vamos a proponer para este estudio, vamos a observar como diferentes variables, pueden afectar o no a las tasas de criminalidad entendiendo dicho factor como el costo del castigo aplicado una vez se comete un delito. Nuestra variable objetivo será Crime que es la tasa de criminalidad y está referenciada como número de delitos por cada 100.000 habitantes en 1960.

Para estudiar qué factores pueden estar asociados a las diferencias en las tasas de criminalidad entre los estados, vamos a llevar a cabo un análisis computacional basado en un conjunto de regresiones lineales. El propósito de este estudio inicial, es explorar cómo se comporta cada variable del US Crime Dataset cuando se analiza de manera independiente como posible determinante de la criminalidad. Este enfoque nos permite obtener una visión preliminar del papel que podrían desempeñar diferentes factores económicos, sociales y policiales en la explicación del fenómeno delictivo.

El conjunto de datos que hemos utilizado recoge información diversa relacionada con el nivel de riqueza, el gasto policial, la estructura demográfica, el desempleo, la desigualdad, la probabilidad de arresto y otros elementos que la literatura suele considerar importantes a la hora de analizar el comportamiento delictivo. Al estimar modelos en los que la criminalidad actúa como variable dependiente y cada predictor se incorpora por separado, es posible identificar patrones de asociación, evaluar su relevancia estadística y estimar la proporción de variabilidad explicada por cada variable.

Aunque el análisis de una sola variable no capta la complejidad real del fenómeno, dado que la criminalidad es el resultado de múltiples factores que interactúan entre sí, sí ofrece un punto de partida sólido para comprender la relevancia individual de cada variable antes de avanzar hacia modelos más completos. Esta primera parte, ofrecerá una panorámica inicial del comportamiento de cada predictor, preparando el terreno para el posterior análisis multivariable.

Tras realizar las regresiones lineales simples y comprobar qué variables presentan una relación estadísticamente significativa con la tasa de criminalidad, es necesario seleccionar aquellas que han demostrado alcanzar una mayor relevancia. Con este conjunto de predictores se elaborará posteriormente un modelo de regresión lineal múltiple.

La finalidad de este modelo más completo es examinar cómo influyen estas variables cuando actúan de forma conjunta. La criminalidad es un fenómeno complejo que no puede explicarse a partir de un único elemento, sino que surge de la interacción de múltiples factores de carácter económico, social o policial. Mientras que las regresiones individuales permiten detectar indicios iniciales de relación, la regresión múltiple ofrece la posibilidad de analizar el impacto específico de cada variable teniendo en cuenta la presencia de las demás, lo que proporciona una interpretación más ajustada a la realidad.

La regresión lineal múltiple permitirá identificar qué variables continúan siendo relevantes cuando se consideran dentro de un mismo modelo, valorar su importancia relativa y determinar el grado en que explican las diferencias en las tasas de criminalidad entre los estados. Este paso es esencial para obtener resultados más consistentes y respaldados por un análisis estadístico riguroso.

METODOLOGÍA

El análisis se ha desarrollado utilizando el lenguaje de programación R, se trata de un leguaje estadística bastante utilizado debido a su capacidad para manejar datos, generar modelos y ofrecer resultados que se puedan reproducir con posterior . Para el estudio computacional vamos a utilizar el software R-Studio y utilizaremos el método que continuación describiremos.

En primer lugar analizaremos el conjunto de datos con el que vamos a trabajar (US Crime Dataset) y seleccionaremos las variables necesarias para el análisis. Posteriormente, revisaremos los datos para asegurar su correcta estructura y dividiremos el conjunto en datos de entrenamiento para poder realizar predicciones posteriores.

Para realizar el estudio computacional, vamos a hacer una división del conjunto de datos con una proporción de 80/20 utilizando como conjunto de entrenamiento un total de 38 estados y como conjunto prueba 9 estados. El conjunto total consta de 47 estados.

Como R nos permitirá calcular los coeficientes estimados, los p-valores y el coeficiente de determinación (r^2), que utilizaremos para evaluar la relevancia estadística y la capacidad explicativa de cada modelo. Utilizando de variable objetivo como ya hemos dicho antes, la variable Crime, crearemos diferentes modelos comparando las variables de conjunto de datos, dos a dos.

Interpretaremos los resultados para cada modelo analizando su relevancia estadística y la proporción de variabilidad explicada por la variable. Esto nos permitirá comparar el comportamiento de los distintos predictores.

Finalmente, elaboraremos una tabla comparativa. Construiremos una tabla resumen con los valores de r^2, la significancia estadística y una breve interpretación de cada variable, con el objetivo de sintetizar los resultados de manera clara y comparable.

El uso de un enfoque computacional nos permitirá realizar el análisis de forma sistemática y replicable, garantizando que los resultados puedan reproducirse y ampliarse en fases posteriores del trabajo, especialmente en el análisis con más de una variable.

3.3.Presentación de resultados regresión.

REGRESIÓN LINEAL SIMPLE

MODELO 1 REGRESIÓN LINEAL SIMPLE (RLS). VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'M'.

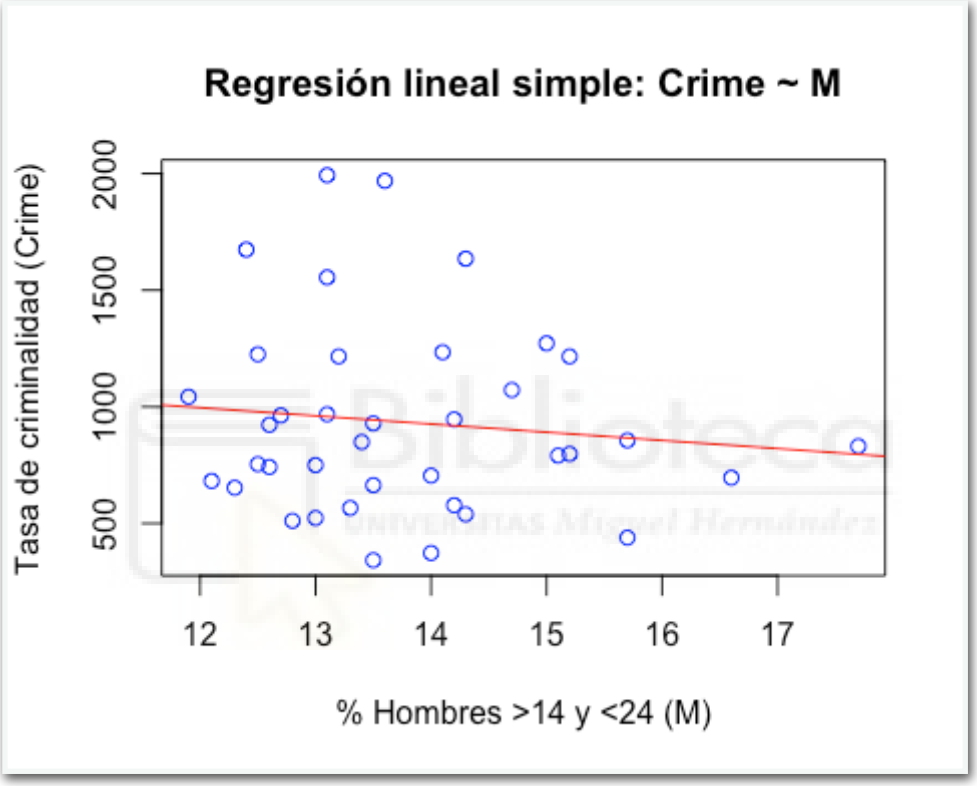
Con este primer modelo, intentaremos predecir si el porcentaje de hombres jóvenes (M) tiene relación con la tasa de criminalidad (Crime).

```
regresionM <- lm(Crime ~ M, datosEntrenamiento)
summary(regresionM)

##
## Call:
## lm(formula = Crime ~ M, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -601.46 -302.92  -68.88  227.45 1035.46
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1418.57     728.09   1.948   0.0592 .
## M             -35.19      52.61  -0.669   0.5078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 414.6 on 36 degrees of freedom
## Multiple R-squared:  0.01228,    Adjusted R-squared:  -0.01516
## F-statistic: 0.4476 on 1 and 36 DF,  p-value: 0.5078
```

En el modelo de regresión lineal simple obtenido, el valor del coeficiente de determinación (r^2) es de 0.01228, lo que indica que únicamente el 1,2 % de la variabilidad de la tasa de criminalidad (Crime) es explicada por la variable

independiente M (porcentaje de hombres jóvenes). Este resultado sugiere que el modelo no logra explicar una relación significativa entre ambas variables y que la variación observada en la tasa de criminalidad se debe, en un 98,8 %, a otros factores no incluidos en el modelo o al componente aleatorio. En consecuencia, se concluye que la variable M, por sí sola, no constituye un buen predictor de la criminalidad en este conjunto de datos.



MODELO 2 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'ED'

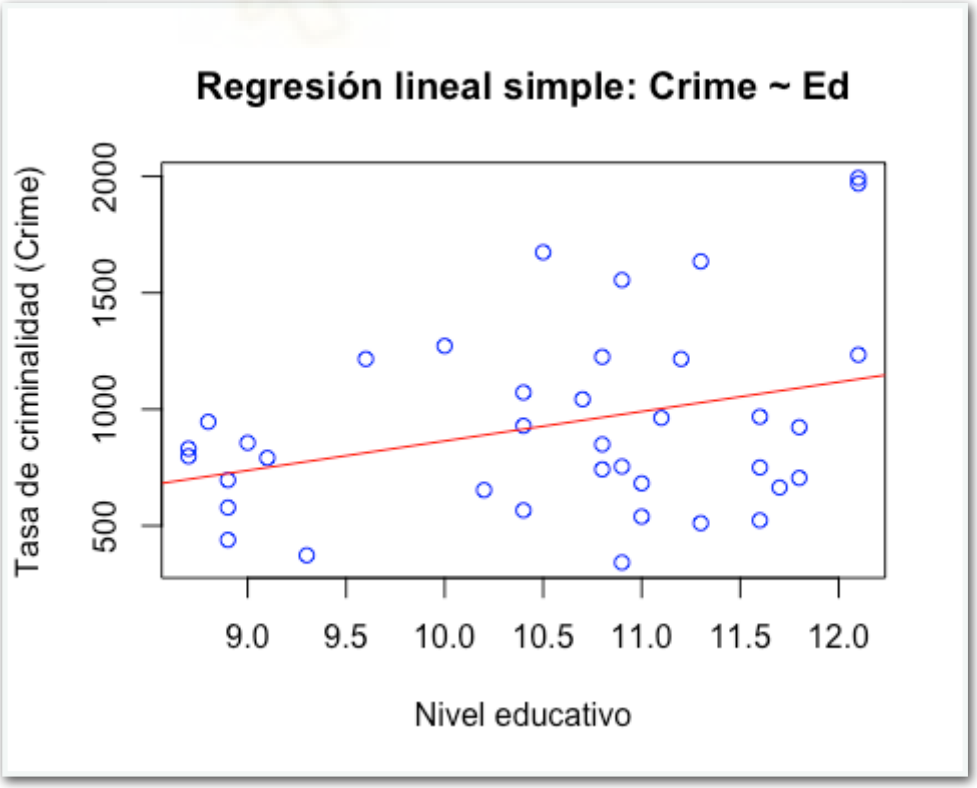
Este modelo analiza si existe una relación estadísticamente significativa entre el nivel educativo medio y la tasa de criminalidad en los estados de EEUU.

```
regresionEd <- lm(Crime ~ Ed, datosEntrenamiento)
summary(regresionEd)
```

```
##
## Call:
```

```
## lm(formula = Crime ~ Ed, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -635.85 -302.81  -34.44  189.49  863.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -401.79     633.66  -0.634   0.5300
## Ed            126.57      59.76   2.118   0.0411 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 393.4 on 36 degrees of freedom
## Multiple R-squared:  0.1108, Adjusted R-squared:  0.08611
## F-statistic: 4.486 on 1 and 36 DF,  p-value: 0.04114
```

El coeficiente de determinación ($r^2 = 0.1108$) muestra que el modelo explica aproximadamente el 11,1 % de la variabilidad de la tasa de criminalidad. Aunque el modelo sugiere una relación significativa entre educación y criminalidad, la baja capacidad explicativa (r^2 bajo) pone de manifiesto que la tasa de criminalidad depende de otros factores socioeconómicos y demográficos adicionales



MODELO 3 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'Po1'

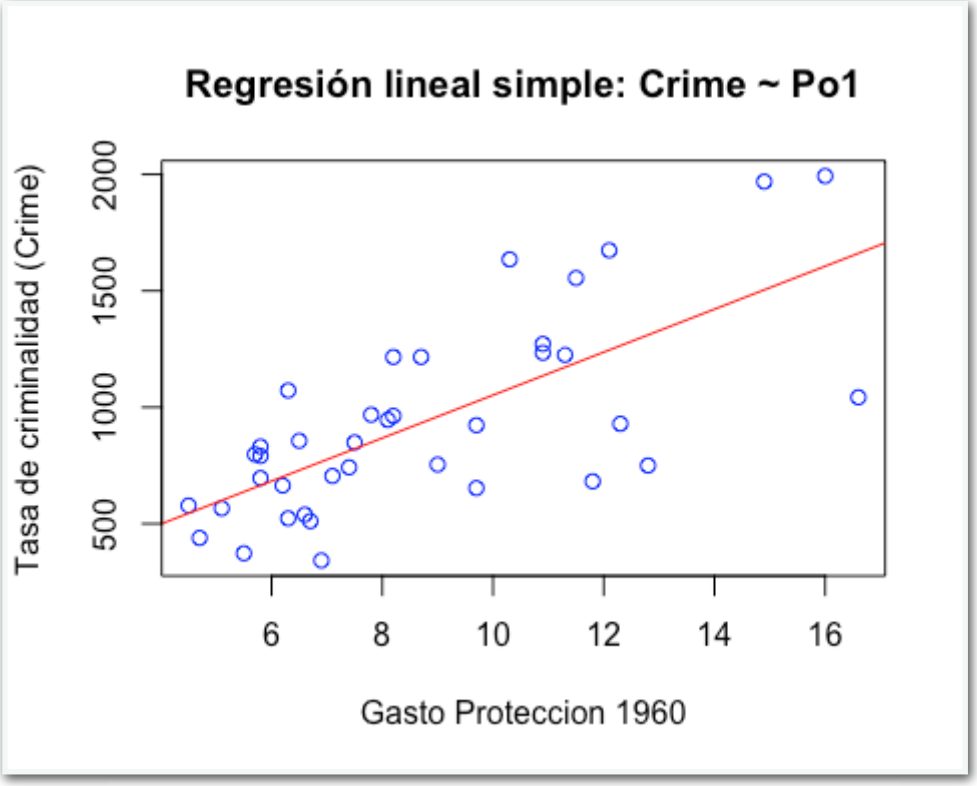
Este modelo analiza si existe una relación entre el gasto policial en 1959 (Po1) y la tasa de criminalidad (Crime). Intentaremos responder a la pregunta "¿Aumentar el gasto policial está asociado con una disminución/aumento de la criminalidad?"

```
regresionPo1 <- lm(Crime ~ Po1, datosEntrenamiento)
summary(regresionPo1)
```

```
##
## Call:
## lm(formula = Crime ~ Po1, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -617.73 -196.49   32.33  141.20  555.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   129.68     142.84   0.908     0.37
## Po1            92.23      15.44   5.975 7.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 295.6 on 36 degrees of freedom
## Multiple R-squared:  0.4979, Adjusted R-squared:  0.4839
## F-statistic:  35.7 on 1 and 36 DF,  p-value: 7.514e-07
```

El valor del coeficiente de determinación (r^2 = 0.4979) muestra que el modelo explica aproximadamente el 49,8 % de la variabilidad de la tasa de criminalidad a partir del gasto policial. En conjunto, los resultados reflejan una asociación estadísticamente significativa y relevante entre el gasto policial y la criminalidad.

Este resultado puede interpretarse como una relación de respuesta institucional, es decir, los estados con mayores niveles de criminalidad tienden a destinar también mayores recursos a seguridad pública.



MODELO 4 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'Po2'

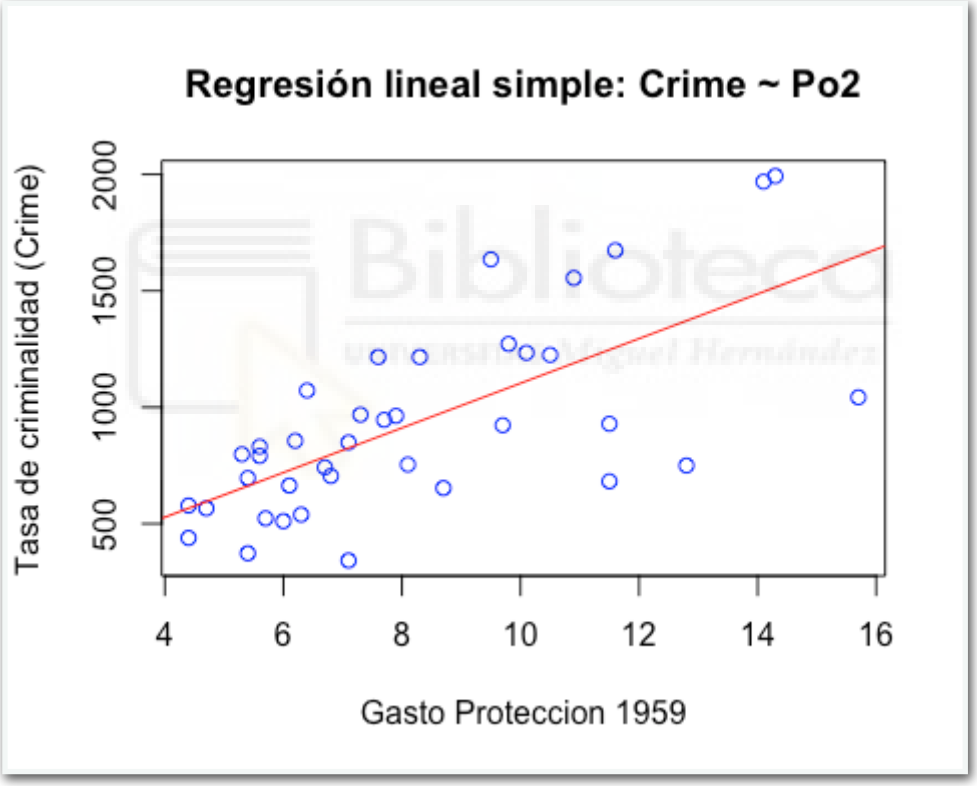
Este modelo analiza si existe una relación entre el gasto policial en 1960 (Po2) y la tasa de criminalidad (Crime). Como en el modelo anterior (modelo 3), intentaremos responder a la pregunta "¿Aumentar el gasto policial está asociado con una disminución/aumento de la criminalidad?"

```
regresionPo2 <- lm(Crime ~ Po2, datosEntrenamiento)
summary(regresionPo2)
```

```
##
## Call:
## lm(formula = Crime ~ Po2, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -621.57 -167.60   28.94  148.70  579.83
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   144.31     148.69   0.971    0.338
```

```
## Po2           95.88      17.04   5.627 2.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 304.3 on 36 degrees of freedom
## Multiple R-squared:  0.468,  Adjusted R-squared:  0.4532
## F-statistic: 31.66 on 1 and 36 DF,  p-value: 2.19e-06
```

El coeficiente de determinación ($r^2 = 0.468$) indica que el modelo explica aproximadamente el 46,8 % de la variabilidad de la tasa de criminalidad a partir del gasto policial en 1960. Este valor refleja una capacidad explicativa considerable para tratarse de un modelo con una sola variable independiente.



COMPARATIVA GASTO POLICIAL

1959 (Po1) y GASTO POLICIAL 1960 (Po2)

Al comparar los modelos de regresión lineal simple que utilizan el gasto policial en 1959 (Po1) y en 1960 (Po2) como variables explicativas de la tasa de criminalidad, se observan resultados coherentes y de magnitud similar. En ambos casos, los coeficientes estimados son positivos y altamente significativos, lo que

indica que los estados con mayores niveles de criminalidad presentan también un mayor gasto policial.

Este patrón sugiere una relación de respuesta institucional, en la que el incremento del presupuesto destinado a seguridad pública no necesariamente reduce la criminalidad, sino que puede estar motivado por niveles delictivos ya elevados.

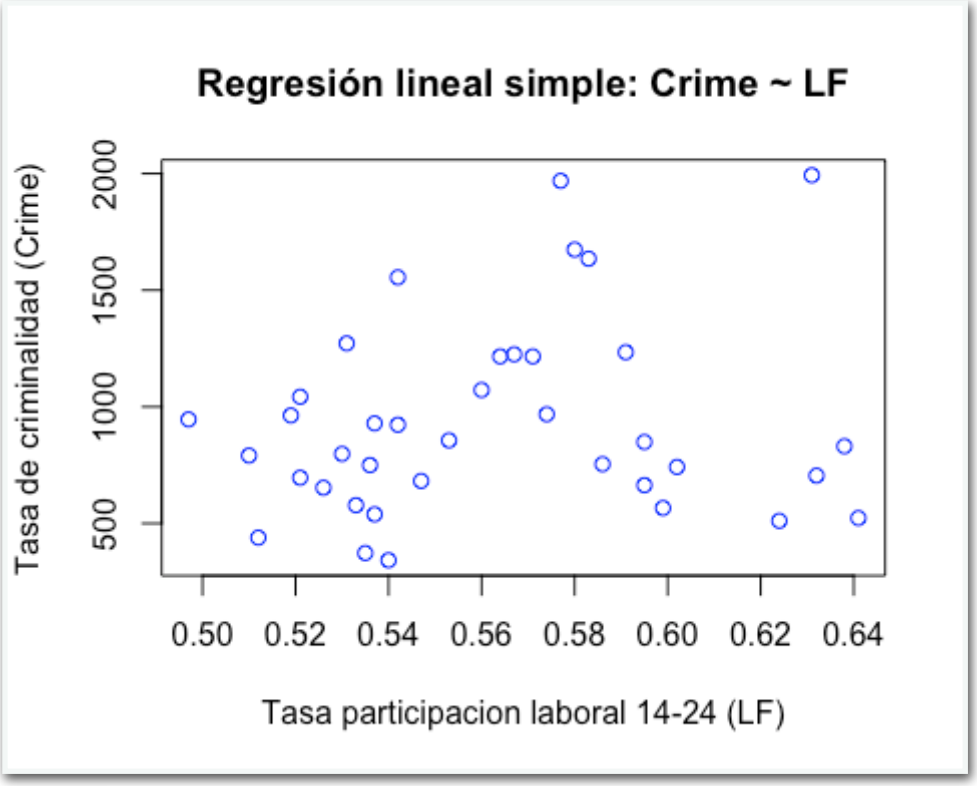
MODELO 5 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'LF'

Este modelo analiza el efecto de la tasa de participación laboral de homrbres entre 14 y 24 años (LF) sobre la criminalidad, evaluando si el nivel de actividad económica de la población está asociado con mayores o menores niveles de delincuencia.

```
regresionLF <- lm(Crime ~ LF, datosEntrenamiento)
summary(regresionLF)
```

```
##
## Call:
## lm(formula = Crime ~ LF, data = datosEntrenamiento)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -553.4 -293.7  -70.9  234.3 1011.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -16.84     978.44  -0.017    0.986
## LF           1689.26    1735.06   0.974    0.337
##
## Residual standard error: 411.8 on 36 degrees of freedom
## Multiple R-squared:  0.02566,    Adjusted R-squared:  -0.00141
## F-statistic: 0.9479 on 1 and 36 DF,  p-value: 0.3367
```

El valor del coeficiente de determinación (r^2 = 0.02566) muestra que el modelo explica únicamente el 2,5 % de la variabilidad de la tasa de criminalidad. Esto implica que la variable LF no aporta capacidad explicativa y que casi toda la variación en la criminalidad se debe a otros factores no incluidos en este modelo simple.



MODELO 6 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'Pop'

Este modelo evalúa si el tamaño de la población de cada estado está asociado con su nivel de criminalidad. La pregunta que intentaremos responder es ¿Los estados con mayor población presentan mayores tasas de criminalidad?

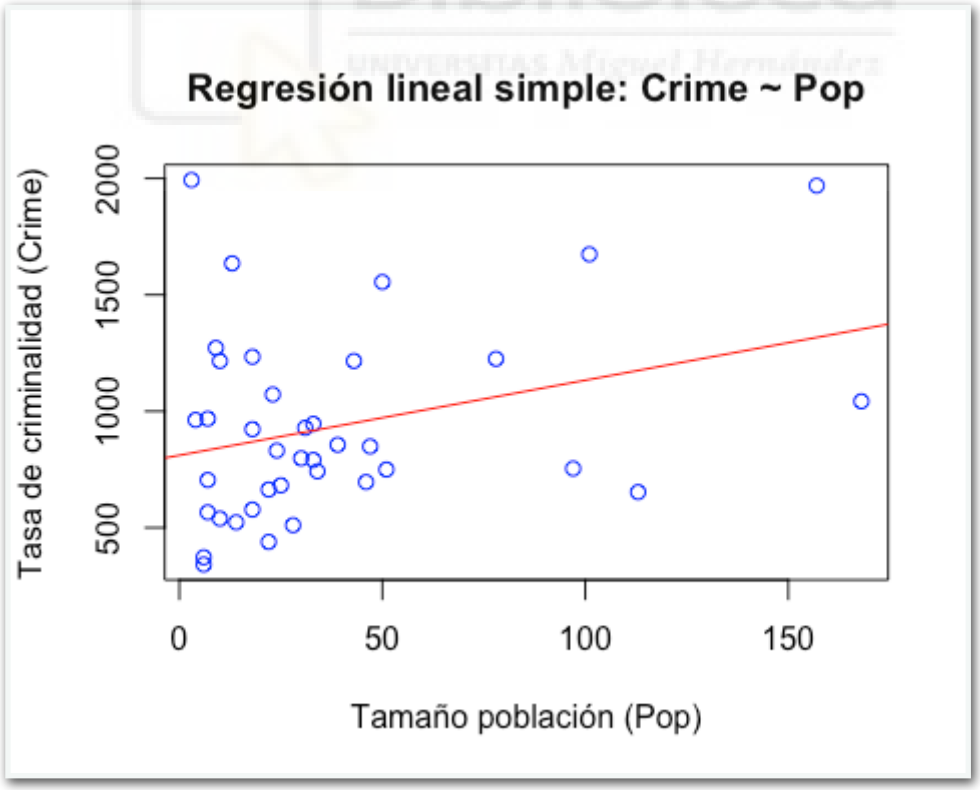
```
regresionPop <- lm(Crime ~ Pop, datosEntrenamiento)
summary(regresionPop)
```

```
##
## Call:
## lm(formula = Crime ~ Pop, data = datosEntrenamiento)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -522.4 -285.1 -111.6  180.8 1172.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  810.905     89.003   9.111 7.03e-11 ***
## Pop            3.225      1.620   1.991   0.0541 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 395.9 on 36 degrees of freedom
## Multiple R-squared:  0.09916,    Adjusted R-squared:  0.07414
## F-statistic: 3.963 on 1 and 36 DF,  p-value: 0.05415
```

El coeficiente de determinación (r^2 = 0.09916) muestra que el modelo explica aproximadamente el 9,9 % de la variabilidad de la tasa de criminalidad a partir del tamaño de la población. Esto implica que la variable Pop posee una capacidad explicativa limitada, y que la mayor parte de la variación en los niveles de criminalidad depende de otros factores.

En conjunto, los resultados sugieren que, aunque pudiera existir una ligera relación entre población y criminalidad, el tamaño poblacional por sí solo no constituye un predictor sólido de la tasa delictiva en los datos analizados.



MODELO 7 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'NW'

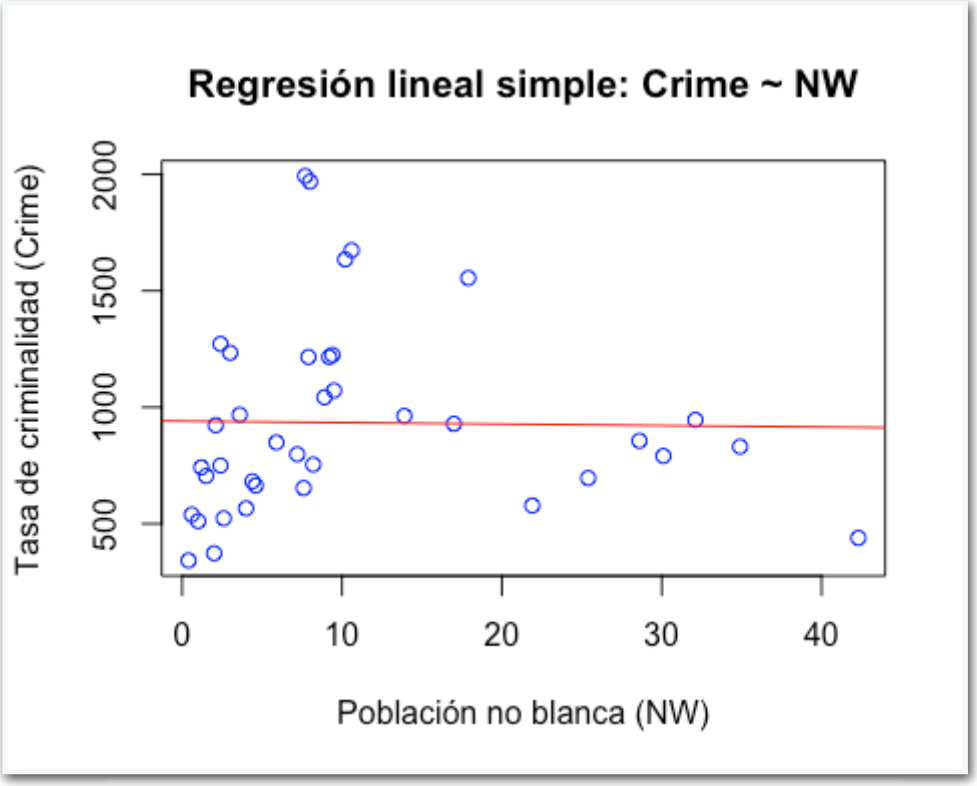
Este modelo estudia si el porcentaje de población no blanca (NW) está asociado con los niveles de criminalidad (Crime) entre los distintos estados de Estados Unidos

```
regresionNW <- lm(Crime ~ NW, datosEntrenamiento)
summary(regresionNW)
```

```
##
## Call:
## lm(formula = Crime ~ NW, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -598.39 -269.16  -87.23  244.81 1057.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  940.655     96.245   9.774 1.14e-11 ***
## NW            -0.658      6.340  -0.104    0.918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 417.1 on 36 degrees of freedom
## Multiple R-squared:  0.000299,   Adjusted R-squared:  -0.02747
## F-statistic: 0.01077 on 1 and 36 DF,  p-value: 0.9179
```

El coeficiente de determinación (r^2 = 0.000299) muestra que el modelo explica tan solo el 0,03 % de la variabilidad de la tasa de criminalidad. Este valor es prácticamente nulo, lo que implica que la variable NW no aporta capacidad explicativa alguna en este modelo.

En otras palabras, la criminalidad es independiente del porcentaje de población no blanca cuando se analiza esta variable de forma aislada en el conjunto de datos. En conjunto, los resultados indican que NW no constituye un predictor relevante de la tasa de criminalidad en este conjunto de datos, reforzando la idea de que la criminalidad depende de otros factores socioeconómicos y estructurales mucho más influyentes que la composición racial de la población.



MODELO 8 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'U1'

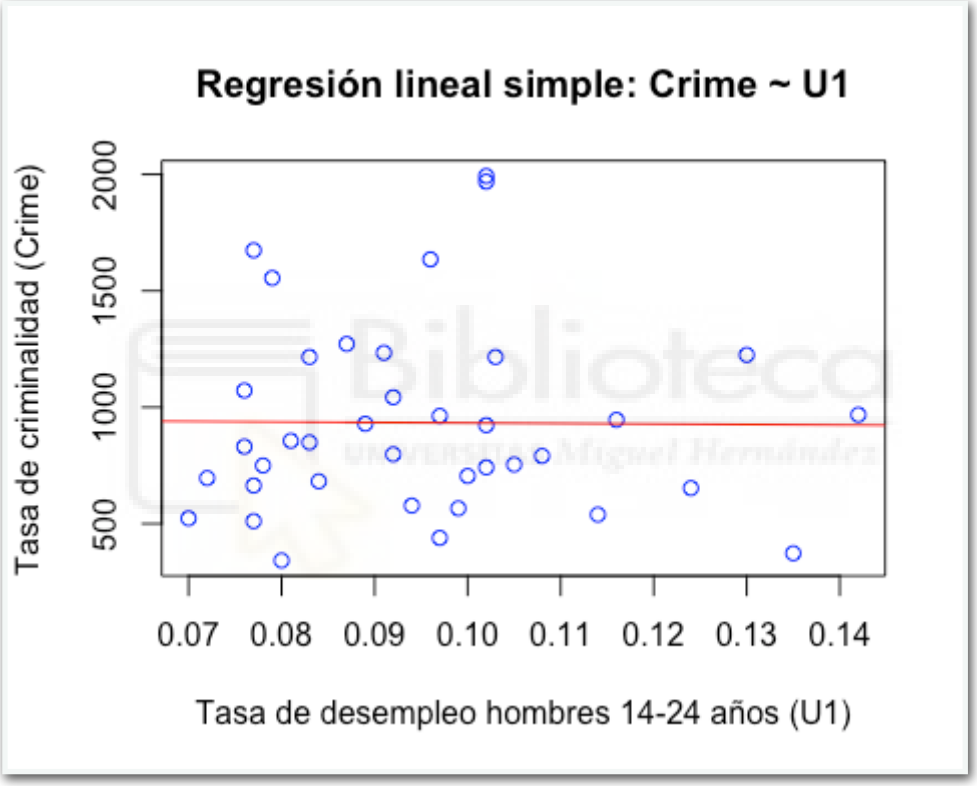
Este modelo analiza si el desempleo entre hombres de 14 a 24 años (U1) es un factor que explica las diferencias en las tasas de criminalidad entre los estados.

```
regresionU1 <- lm(Crime ~ U1, datosEntrenamiento)
summary(regresionU1)
```

```
##
## Call:
## lm(formula = Crime ~ U1, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -595.05 -268.84  -97.17  243.24 1061.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    955.6      374.1   2.554    0.015 *
## U1            -232.4     3870.8  -0.060    0.952
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 417.2 on 36 degrees of freedom
## Multiple R-squared:  0.0001001,  Adjusted R-squared:  -0.02767
## F-statistic: 0.003605 on 1 and 36 DF,  p-value: 0.9525
```

El coeficiente de determinación (r^2 = 0.0001001) muestra que el modelo explica únicamente el 0,01 % de la variabilidad en la criminalidad. Esto implica que la variable U1 no aporta capacidad explicativa alguna.



MODELO 9 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'U2'

Este modelo analiza si el desempleo entre hombres de 35 a 39 años (U2) es un factor que explica las diferencias en las tasas de criminalidad entre los estados.

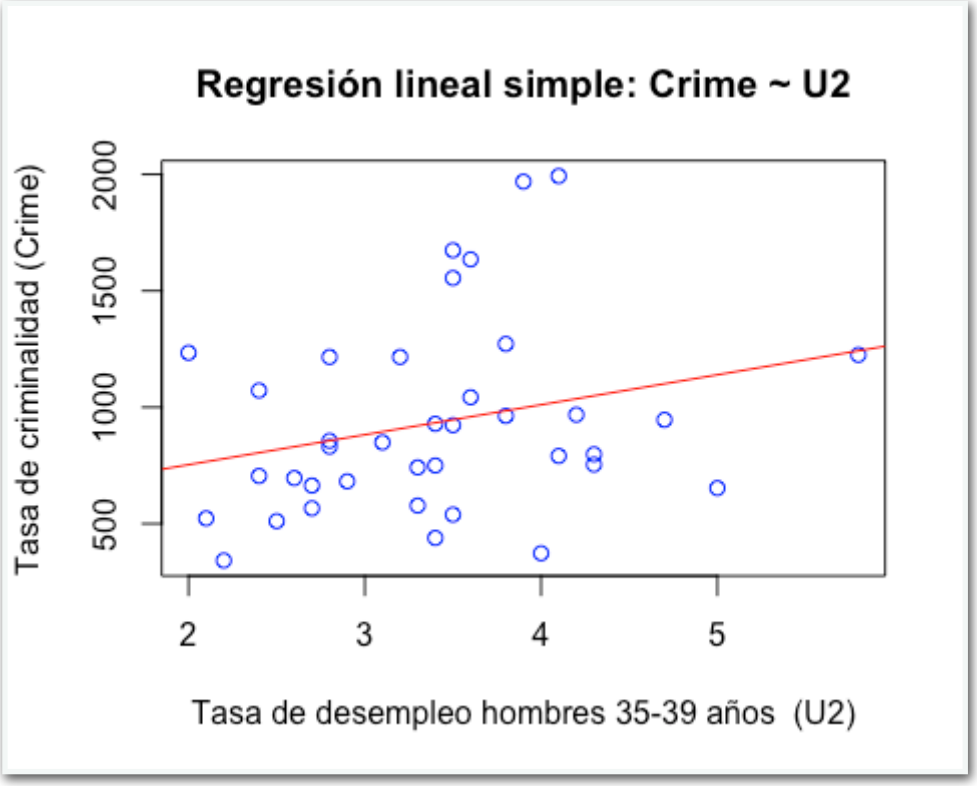
```
regresionU2 <- lm(Crime ~ U2, datosEntrenamiento)
summary(regresionU2)
```

```
##
## Call:
```

```
## lm(formula = Crime ~ U2, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -637.78 -249.35  -84.18  221.29  971.09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   495.95     280.13   1.770   0.0851 .
## U2            128.71      80.12   1.607   0.1169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 403 on 36 degrees of freedom
## Multiple R-squared:  0.06689,    Adjusted R-squared:  0.04098
## F-statistic: 2.581 on 1 and 36 DF,  p-value: 0.1169
```

El coeficiente de determinación ($r^2 = 0.06689$) muestra que el modelo solo logra explicar el 6,7 % de la variabilidad en la tasa de criminalidad, lo que evidencia que la variable U2 tiene una capacidad explicativa muy limitada.

En definitiva, los resultados indican que el desempleo entre hombres de 35 a 39 años no es un predictor relevante de la criminalidad en este conjunto de datos.



COMPARATIVA TASA DESEMPLEO

HOMBRES 14-24 (U1) y HOMBRES 35-39 (U2)

En conjunto, ambos modelos indican que ninguna de las dos tasas de desempleo, ni la correspondiente a hombres jóvenes ni la de adultos de 35 a 39 años, constituye un predictor sólido de la criminalidad. Aunque U2 presenta una relación algo más marcada que U1, ambos modelos muestran falta de significancia estadística y valores de r^2 muy reducidos, evidenciando que la criminalidad depende principalmente de otros factores socioeconómicos no incluidos en estas regresiones simples.

MODELO 11 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'Wealth'

Este modelo intenta responder a la pregunta ¿Influye el nivel de riqueza (Wealth) en las diferencias de criminalidad entre los estados?

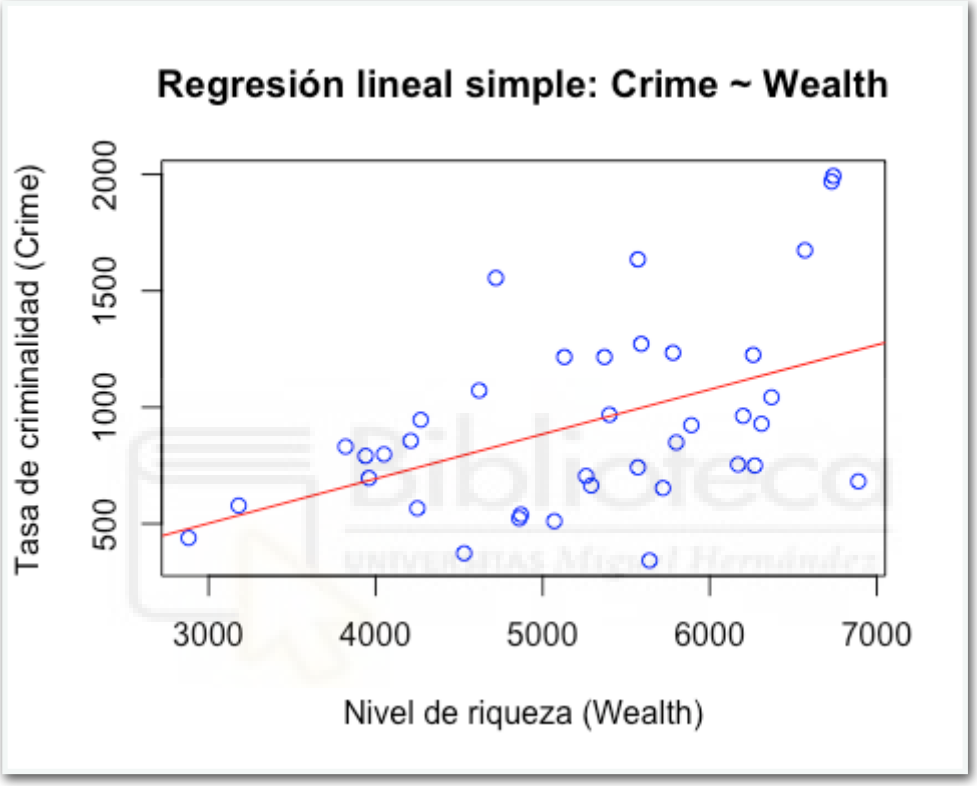
```
regresionWealth <- lm(Crime ~ Wealth, datosEntrenamiento)
summary(regresionWealth)
```

```
##
## Call:
## lm(formula = Crime ~ Wealth, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -664.98 -269.86  -16.21  201.09  775.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -73.15879  318.77391  -0.230  0.81978
## Wealth        0.19151    0.05957   3.215  0.00275 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 367.7 on 36 degrees of freedom
## Multiple R-squared:  0.2231, Adjusted R-squared:  0.2015
## F-statistic: 10.34 on 1 and 36 DF,  p-value: 0.002755
```

El coeficiente de determinación ($r^2 = 0.2231$) muestra que el modelo explica aproximadamente el 22,3 % de la variabilidad de la tasa de criminalidad a partir del nivel de riqueza. Se trata de una capacidad explicativa moderada, claramente

superior a la observada en variables como las que ya hemos analizado, U1, U2, NW o LF, aunque todavía insuficiente para describir por completo el fenómeno.

En conjunto, los resultados señalan que Wealth es un predictor estadísticamente significativo y con cierta capacidad explicativa de la criminalidad en este conjunto de datos.



MODELO 12 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'Ineq'

Este modelo analiza si la desigualdad económica es un factor relacionado con los niveles de criminalidad en los estados de Estados Unidos. Respondería a la pregunta ¿Influye la desigualdad económica en la criminalidad?

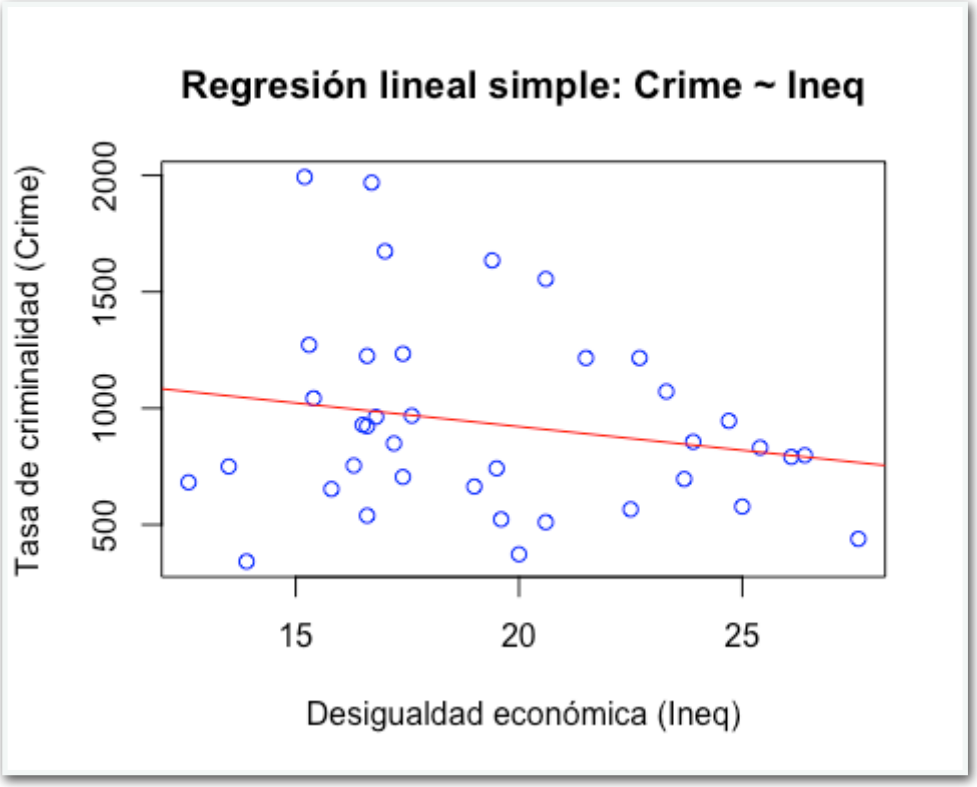
```
regresionIneq <- lm(Crime ~ Ineq, datosEntrenamiento)
summary(regresionIneq)
```

```
##
## Call:
## lm(formula = Crime ~ Ineq, data = datosEntrenamiento)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -702.86 -296.50  -42.86  230.98  981.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1327.91     332.45   3.994 0.000306 ***
## Ineq          -20.36      16.82  -1.211 0.233944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 408.9 on 36 degrees of freedom
## Multiple R-squared:  0.03912,    Adjusted R-squared:  0.01243
## F-statistic: 1.466 on 1 and 36 DF,  p-value: 0.2339
```

El coeficiente de determinación ($r^2 = 0.03912$) indica que el modelo solo consigue explicar el 3,9 % de la variabilidad en la tasa de criminalidad. Este porcentaje tan reducido evidencia que la variable Ineq posee una capacidad explicativa muy limitada, y que la práctica totalidad de las diferencias en los niveles de criminalidad entre estados se debe a otro factores.

En consecuencia, los resultados muestran que la desigualdad económica (Ineq) no actúa como un predictor relevante de la criminalidad en este conjunto de datos.



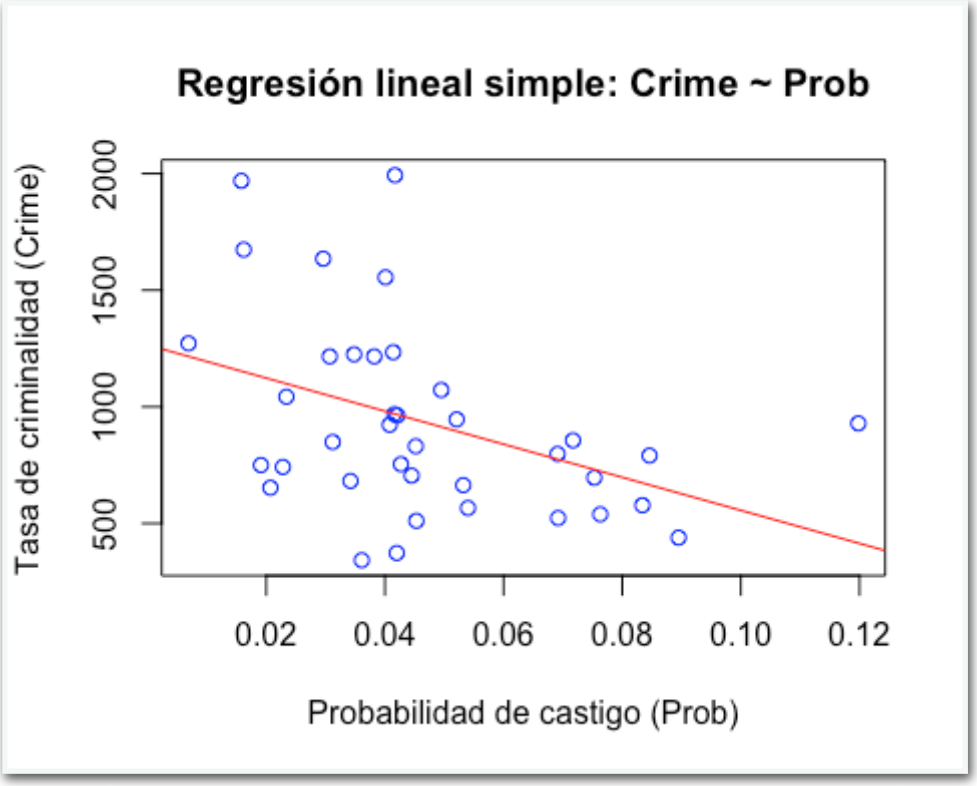
MODELO 13 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'Prob'

Este modelo estudia si la probabilidad de que un delincuente sea arrestado o condenado influye en la criminalidad. Intenta responder a la pregunta, ¿Disminuye la criminalidad cuando aumenta la probabilidad de arresto o castigo?

```
regresionProb <- lm(Crime ~ Prob, datosEntrenamiento)
summary(regresionProb)
```

```
##
## Call:
## lm(formula = Crime ~ Prob, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -666.82 -239.03  -43.56  166.22 1023.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1265.0      137.4   9.204 5.43e-11 ***
## Prob         -7096.3     2629.4  -2.699   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 380.5 on 36 degrees of freedom
## Multiple R-squared:  0.1683, Adjusted R-squared:  0.1452
## F-statistic: 7.284 on 1 and 36 DF,  p-value: 0.01053
```

El coeficiente de determinación ($r^2 = 0.1683$) indica que el modelo explica aproximadamente el 16,8 % de la variabilidad de la tasa de criminalidad a partir de la probabilidad de arresto. Aunque este porcentaje no es muy elevado, sí refleja una capacidad explicativa lo que sugiere que la probabilidad de arresto desempeña un papel más relevante en la explicación del crimen.



MODELO 14 RLS. VARIABLE INDEPENDIENTE 'Crime', VARIABLE DEPENDIENTE 'Time'

En la teoría económica del crimen un aumento en la severidad del castigo, en este caso, condenas más largas, debería reducir la criminalidad por su efecto disuasorio. Con este modelo vamos a intentar responder a la pregunta ¿Influye la duración de la condena en la tasa de criminalidad?

```
regresionTime <- lm(Crime ~ Time, datosEntrenamiento)
summary(regresionTime)
```

```
##
## Call:
## lm(formula = Crime ~ Time, data = datosEntrenamiento)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -601.1 -254.0 -119.6  212.4 1089.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  768.809    272.403   2.822  0.00772 **
## Time           6.114      9.796   0.624  0.53647
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 414.9 on 36 degrees of freedom
## Multiple R-squared:  0.0107, Adjusted R-squared:  -0.01678
## F-statistic: 0.3895 on 1 and 36 DF,  p-value: 0.5365
```

El coeficiente de determinación ($r^2 = 0.0107$) muestra que el modelo explica únicamente el 1,07 % de la variabilidad de la tasa de criminalidad. En conjunto, los resultados indican que la duración media de la condena no constituye un predictor válido de la criminalidad en este conjunto de datos.

La baja significancia estadística y el escaso porcentaje de variación explicada resaltan la necesidad de incorporar otros factores socioeconómicos y estructurales mediante modelos multivariantes para comprender mejor los determinantes del crimen.

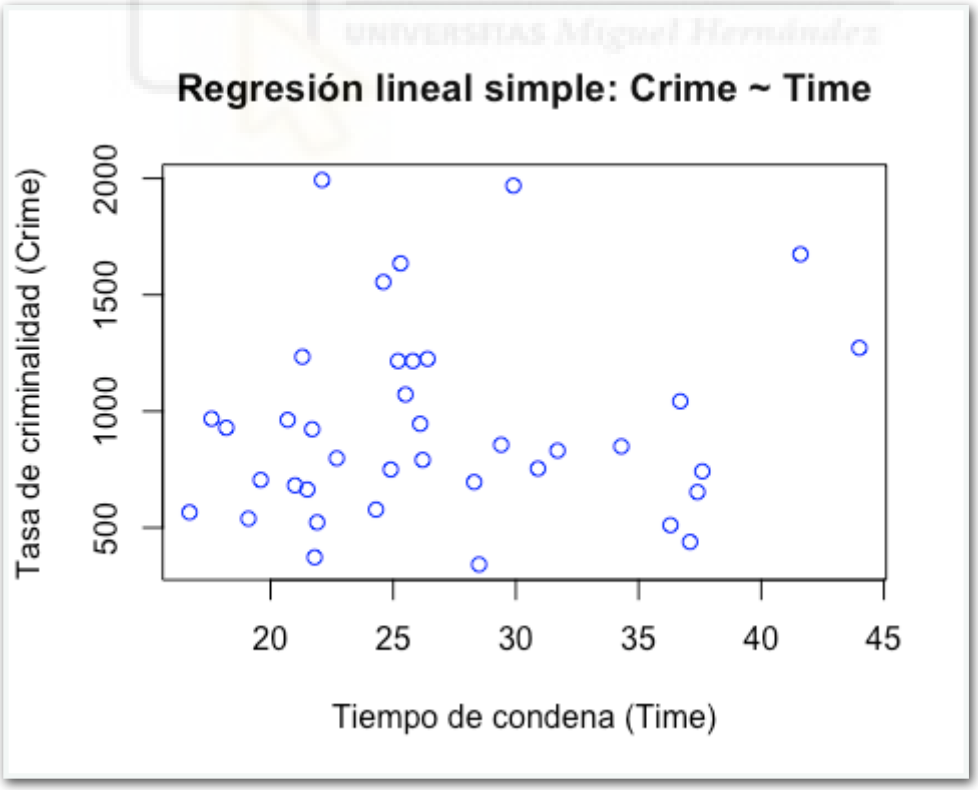


TABLA COMPARATIVA RESULTADOS OBTENIDOS EN ANÁLISIS REGRESIÓN LINEAL SIMPLE

Variable predictora	r^2	Relevancia	Interpretación
M	0.01	No	No explica la criminalidad.
Ed	0.11	Sí	Relación positiva con explicación limitada.
Po1	0.50	Sí	Alta capacidad explicativa; relación positiva.
Po2	0.47	Sí	Alta capacidad explicativa; relación positiva.
LF	0.03	No	No existe relación significativa.
Pop	0.10	Límite	Relación débil y poco concluyente.
NW	0.00	No	No influye en la criminalidad.
U1	0.00	No	No predice el crimen.
U2	0.07	No	Capacidad explicativa muy baja.
Wealth	0.22	Sí	Relación positiva y explicación baja.
Ineq	0.04	No	No aporta relevancia explicativa.
Prob	0.17	Sí	Efecto disuasorio
Time	0.01	No	No explica la criminalidad.

REGRESIÓN LINEAL MÚLTIPLE

MODELO 1 RLM. VARIABLE INDEPENDIENTE ‘CRIME’, VARIABLES DEPENDIENTES ‘Po1’, ‘Po2’, ‘Wealth’, ‘Prob’

Para evaluar la influencia de distintos factores socioeconómicos y policiales sobre la tasa de criminalidad, se ha estimado un modelo de regresión lineal múltiple utilizando los datos del conjunto US Crime. La variable dependiente sigue siendo Crime, mientras que como predictores se han incluido Po1, Po2, Wealth y Prob. Estas variables representan dimensiones relevantes como el gasto policial en 1960 y 1959 respectivamente, el indicador de riqueza y la probabilidad de arresto o condena.

La selección de estos predictores se basó en un análisis previo del comportamiento de los modelos de regresión lineal simple, donde se estudiaron las relaciones “dos a dos” entre Crime y cada una de las variables explicativas por separado. A partir de ese análisis inicial se ha identificado que Po1, Po2, Wealth y Prob eran las variables que mostraban una asociación más significativa con la tasa de criminalidad, por lo que se ha considerado pertinente incorporarlas de manera conjunta en un modelo múltiple para profundizar en su influencia.

El modelo de regresión lineal múltiple permite cuantificar el efecto individual de cada variable controlando el resto, lo que facilita identificar cuáles mantienen un papel relevante cuando se analizan conjuntamente. Asimismo, proporciona información sobre la dirección y magnitud de cada relación, así como sobre la capacidad global del modelo para explicar la variabilidad observada en los niveles de criminalidad entre estados.

El propósito de este análisis es determinar qué factores contribuyen de forma más consistente a explicar las diferencias en criminalidad y valorar si ciertos elementos como la disponibilidad de recursos policiales o la probabilidad de detención pueden considerarse predictores sólidos dentro del conjunto de datos utilizado.

```
regresionRLM <- lm(Crime ~ Po1+Po2+Wealth+Prob ,data= datosEntrenamiento)
summary(regresionRLM)
```

```
##
## Call:
## lm(formula = Crime ~ Po1 + Po2 + Wealth + Prob, data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -690.61 -170.29   20.22  149.96  479.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.298e+02  3.869e+02   2.145   0.0394 *
## Po1          3.281e+02  1.461e+02   2.246   0.0315 *
## Po2         -2.294e+02  1.574e+02  -1.458   0.1544
## Wealth       -1.354e-01  8.302e-02  -1.630   0.1125
## Prob        -3.331e+03  2.402e+03  -1.387   0.1748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 284.9 on 33 degrees of freedom
## Multiple R-squared:  0.5726, Adjusted R-squared:  0.5208
## F-statistic: 11.05 on 4 and 33 DF,  p-value: 8.477e-06
```

El modelo obtiene un r^2 de 0.57, lo que significa que las cuatro variables utilizadas (Po1, Po2, Wealth y Prob) son capaces de explicar alrededor del 57% de las diferencias en la tasa de criminalidad entre los estados. Es decir, algo más de la mitad de la variación en el crimen puede entenderse a partir de estos factores.

Cuando se utiliza el r^2 ajustado, que corrige el efecto de añadir varias variables al modelo, el valor baja a 0.52. Esto indica que, aunque el modelo tiene una capacidad explicativa moderada, no todas las variables aportan la misma utilidad cuando se analizan juntas. Aun así, un 52% sigue siendo un porcentaje aceptable en estudios sociales, donde es habitual que intervengan muchos factores difíciles de medir.

En resumen, el r^2 muestra que el modelo explica una parte importante, pero no la totalidad, del comportamiento de la criminalidad. Por lo tanto, estos resultados sugieren que las variables analizadas influyen en el nivel de crimen, pero también que existen otros elementos no incluidos que podrían ayudar a mejorar la explicación del fenómeno.

TABLA COMPARATIVA RESULTADOS OBTENIDOS EN ANÁLISIS REGRESIÓN LINEAL MULTIPLE

Variable predictora	P-value	Relevancia
Po1	0,0315	Sí (p<0,05)
Po2	0,1544	No
Wealth	0,1125	No
Prob	0,1748	No

MODELO 2 RLM. VARIABLE INDEPENDIENTE 'CRIME', VARIABLES DEPENDIENTES 'TODAS LAS VARIABLES'

Para finalizar, vamos a interpretar un modelo de regresión lineal múltiple utilizando como variable independiente Crime y como dependientes el conjunto de todas las demás variables como predictores de la tasa de criminalidad. El objetivo de este modelo es analizar el efecto conjunto de todos los factores socioeconómicos, demográficos y policiales, y determinar cuáles de ellos mantienen un impacto significativo cuando se controlan simultáneamente unos por otros.

```
regresionRLM2 <- lm(Crime ~ . ,data= datosEntrenamiento)
summary(regresionRLM2)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = datosEntrenamiento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -309.13 -104.13  -33.14  116.47  399.91
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7594.8378  2164.2749  -3.509  0.00198 **
## M              74.5493    44.8335   1.663  0.11054
## So            100.5641   174.7852   0.575  0.57089
## Ed            150.1609    70.5539   2.128  0.04476 *
## Po1           182.6350   157.0649   1.163  0.25737
## Po2          -103.8188   174.9515  -0.593  0.55896
## LF         -1109.8071  1693.0149  -0.656  0.51893
## M.F           37.0280    23.1726   1.598  0.12432
## Pop           -0.4669     1.4560  -0.321  0.75147
## NW             2.1385     7.4178   0.288  0.77582
## U1         -6068.4033  4644.5716  -1.307  0.20486
## U2           131.5777    91.6661   1.435  0.16524
## Wealth         0.1801     0.1097   1.641  0.11504
## Ineq          81.9726    24.7287   3.315  0.00315 **
## Prob       -5406.1560  2638.0973  -2.049  0.05255 .
## Time          -0.6905     9.3250  -0.074  0.94164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 205.2 on 22 degrees of freedom
## Multiple R-squared:  0.8521, Adjusted R-squared:  0.7512
## F-statistic: 8.448 on 15 and 22 DF,  p-value: 5.794e-06
```

El modelo estimado incluye todas las variables disponibles en el conjunto de datos como predictores de la tasa de criminalidad (Crime). Globalmente, los resultados muestran un buen ajuste, con un R^2 del 0.8521 y un R^2 ajustado del 0.7512, lo que indica que aproximadamente el 75% de la variabilidad del crimen se explica por el conjunto de variables incluidas.

Teniendo el cuenta el p-valor como indicador de la influencia de las variables sobre Crime, obtenemos la tabla que a continuación presentamos con su correspondiente interpretación.

TABLA COMPARATIVA RLM - CRIME - TODAS LAS VARIABLES

Variable	p-valor	Significación	Interpretación
Ed	0.04476	Significativa (p < 0.05)	Aumentos en educación se asocian con cambios significativos en Crime.
Ineq	0.00315	Muy significativa (p < 0.01)	La desigualdad es un fuerte predictor del crimen.
Prob	0.05255	Casi significativa (p < 0.10)	Tiende a reducir el crimen, relación cercana a significación.
M	0.11054	No significativa	No aporta efecto único en el modelo.
So	0.57089	No significativa	Su relación con Crime no es concluyente.
Po1	0.25737	No significativa	Pierde relevancia al introducir todas las variables.
Po2	0.55896	No significativa	Efecto no concluyente.
LF	0.51893	No significativa	No explica Crime de forma independiente.
M.F	0.12432	No significativa	Efecto incierto.
Pop	0.75147	No significativa	Alta p, no significativa.
NW	0.77582	No significativa	No aporta información relevante.
U1	0.20486	No significativa	No significativo al controlar el resto.
U2	0.16524	No significativa	Sin significación estadística.
Wealth	0.11504	No significativa	Relevante en modelos simples, pero no aquí.
Time	0.94164	No significativa	Totalmente irrelevante en el modelo.

EVALUACIÓN DEL MODELO 2 DE REGRESIÓN LINEAL MULTIPLE

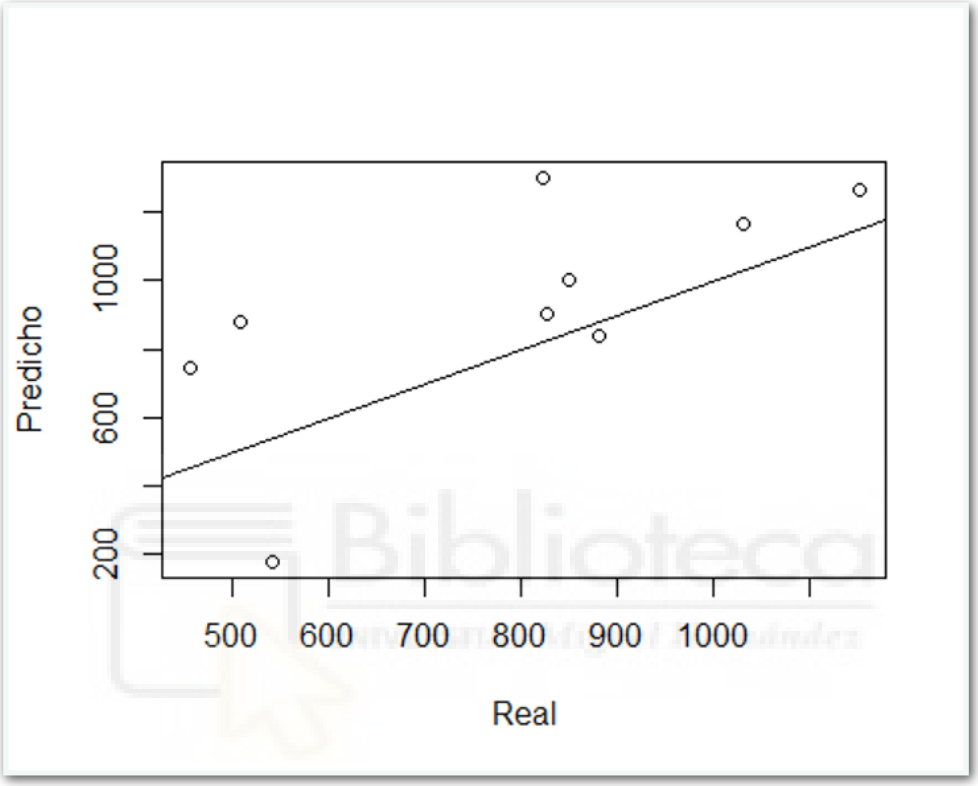
Para finalizar, vamos a realizar la evaluación del modelo de regresión utilizando los datos de prueba, es decir, datos que no se usaron para entrenar el modelo. Esto nos permitirá comprobar si la regresión del modelo 2, es capaz de predecir adecuadamente la criminalidad en observaciones nuevas.

```
predicciones <- predict(regresionRLM2, newdata = datosPrueba)
Error cuadrático medio (MSE)
MSE <- mean( (datosPrueba$Crime - predicciones)^2 )
MSE
## [1] 71578.59
Raíz del error cuadrático medio (RMSE)
RMSE <- sqrt(MSE)
RMSE
## [1] 267.5418
R cuadrado en los datos de prueba
R2 <- 1 - sum((datosPrueba$Crime - predicciones)^2) /
      sum((datosPrueba$Crime - mean(datosPrueba$Crime))^2)
R2
## [1] -0.4161363
```

El MSE (error cuadrático medio) mide, en promedio, cuánto se equivocan las predicciones respecto a los valores reales. Como podemos observar, nuestro MSE es de 71578,59 lo o cual refleja que, aunque existe cierto nivel de error, algo habitual en problemas de predicción, el modelo es capaz de aproximarse razonablemente a los valores reales dentro de la variabilidad propia de la variable Crime.

El RMSE (Raiz del error cuadrático medio) mide cuánto se equivoca un modelo en promedio, pero expresado en las mismas unidades que la variable que estamos prediciendo, en este caso Crime. Como podemos ver, las predicciones del modelo se equivocan en unos 267 puntos. Aunque este valor muestra que existe desviación entre las predicciones y los datos reales, debe interpretarse con prudencia ya que este nivel de error es coherente con la naturaleza del fenómeno y con las limitaciones inherentes a un modelo lineal aplicando datos reales.

En cuanto al r² obtenido en los datos de prueba, su valor negativo indica que el modelo pierde capacidad predictiva cuando se aplica a observaciones nuevas que no formaron parte del entrenamiento. Este comportamiento es totalmente normal y esperable, ya que el modelo se ajusta inicialmente a los datos de entrenamiento.



En la gráfica podemos observar que las predicciones siguen la tendencia general del delito, lo que confirma que el modelo captura parte de la estructura subyacente del fenómeno. La disminución del rendimiento predictivo en los datos de prueba no invalida el modelo, sino que simplemente refleja la complejidad de la variable *Crime* y la necesidad de considerar modelos más sofisticados o conjuntos de datos más amplios en trabajos futuros.

4. DISCUSION

4.1.Interpretación de los resultados.

El conjunto de regresiones lineales simples realizado permite observar cómo distintas variables económicas, demográficas y policiales se relacionan individualmente con la tasa de criminalidad. El análisis evidencia que las capacidades explicativas son muy dispares según la variable considerada.

En las relaciones lineales simples de cada variable con la variable independiente, en primer lugar, podemos destacar las variables Po1 y Po2 (Gastos per cápita en protección policial en 1960 y 1959 respectivamente), la variable Wealth (Valor mediano de los ingresos familiares) y finalmente Prob (Probabilidad de encarcelamiento). Estas son las variables con resultados más robustos dentro de este conjunto de modelos. Tanto Po1 como Po2 (gasto policial en 1960 y 1959) muestran coeficientes significativos y los valores de r^2 más elevados, cercanos al 50 %. Esto significa que por sí solas explican casi la mitad de la variación observada en la criminalidad.

La relación positiva sugiere que los estados con mayores niveles delictivos destinan también mayores recursos policiales, lo que apunta a un posible efecto de respuesta institucional más que a un efecto preventivo.

La variable Wealth también presenta un coeficiente significativo y una capacidad explicativa intermedia baja (22 %). La asociación positiva podría estar vinculada a dinámicas propias de estados con estructuras urbanas y económicas más complejas.

Por su parte, Prob, la probabilidad de arresto, arroja un r^2 cercano al 17 %, lo que respalda su papel como factor disuasorio, cuando la probabilidad de ser detenido aumenta, la criminalidad tiende a reducirse.

En contraste, otras variables presentan resultados mucho menos concluyentes. U1, U2, NW, Ineq, M, LF y Time muestran coeficientes no significativos y valores de r^2 muy bajos, en algunos casos prácticamente nulos. Esto indica que, cuando se analizan de manera aislada, estas variables no ofrecen información relevante para explicar la criminalidad en el conjunto de datos. La explicación de este rendimiento limitado puede deberse a la falta de una relación real o a la influencia de otros factores no considerados en el modelo de regresión lineal simple.

En conjunto, el análisis comparativo pone de manifiesto que ninguna variable aislada es suficiente para comprender la complejidad del fenómeno delictivo. Aunque algunas variables destacan por su mayor capacidad explicativa, la mayoría muestra relaciones débiles o inexistentes. Esto refuerza la idea de que la criminalidad depende de la interacción simultánea de múltiples factores y que, por tanto, resulta necesario avanzar hacia otros modelos que permitan integrar y analizar de manera conjunta las distintas dimensiones que influyen en el comportamiento delictivo.

En segundo lugar, hemos aplicado un modelo de regresión lineal múltiple para comprobar si, al incorporar varias variables de forma conjunta, se obtiene un ajuste mejor y una capacidad explicativa superior.

Al estudiar el modelo de regresión múltiple que incluye las variables más significativas obtenidas en el estudio de regresión lineal simple, se aprecia que Po1 es la única que realmente tiene un efecto claro sobre la tasa de criminalidad, ya que es la única que resulta significativa. Las demás variables no muestran una contribución importante cuando se analizan junto a Po1, por lo que no mejoran el modelo. Por este motivo, los resultados obtenidos con esta regresión lineal múltiple no mejoran sustancialmente la explicación aportada por los modelos de regresión lineal simple anteriormente estudiados. En consecuencia, se ha considerado adecuado explorar un modelo adicional con el fin de evaluar si podía ofrecer un ajuste más apropiado para los datos.

En tercer lugar, utilizando el modelo de regresión múltiple que incluye todas las variables de conjunto de datos y procediendo a su análisis, se observa que

solo un pequeño conjunto de variables mantiene una relación lineal estadísticamente significativa con la criminalidad cuando todas ellas se consideran de forma conjunta.

Destacan el nivel educativo (Ed) y especialmente la desigualdad (Ineq), que se convierte en el predictor más sólido del modelo. La probabilidad de condena o arresto (Prob), también muestra una posible relevancia.

A pesar de que el resto de variables no alcanzan significación estadística, en este enfoque multivariable, el rendimiento global del modelo resulta bastante bueno. De hecho, este último modelo, con un resultado en el coeficiente de determinación de 0,8521, explica el 85% de la variable Crime. Esto muestra que, aun cuando no todas las variables individuales resulten significativas, el modelo en su conjunto ofrece un ajuste adecuado y coherente con la estructura de los datos.

Por ello, se concluye que este modelo de regresión múltiple constituye la mejor opción para estimar la tasa de criminalidad dentro del marco de este estudio, al ofrecer el equilibrio más sólido.

En cuanto a las predicciones efectuadas, la evaluación del modelo sobre el conjunto de prueba indica que su capacidad predictiva es razonable, especialmente teniendo en cuenta que se trabaja con datos reales y con un fenómeno tan complejo como la criminalidad. Tanto el MSE como el RMSE muestran la existencia de cierto error en las predicciones, algo completamente habitual en modelos aplicados a contextos sociales, donde la variabilidad es elevada y difícil de capturar al cien por cien. Aun así, el modelo logra aproximarse a los valores reales y mantener la tendencia general del comportamiento delictivo.

En conjunto, los resultados ponen de manifiesto que, aunque existen limitaciones, como ocurre con cualquier modelo estadístico aplicado a datos reales, la regresión lineal múltiple sigue siendo una herramienta válida capaz de proporcionar estimaciones razonables y coherentes.

4.2.Coherencia de los resultados con estudios previos

Los resultados de este estudio revelan que ciertos factores tradicionalmente vinculados a la criminalidad, como el gasto policial y la desigualdad económica, muestran una relación estadísticamente significativa con la tasa de delitos. Este patrón coincide con los hallazgos de investigaciones clásicas en economía del crimen. Por ejemplo, Ehrlich (1973) argumentaba que las decisiones delictivas pueden verse como elecciones racionales influenciadas por variables económicas y la severidad del castigo. Coincidiendo con esto, en nuestro primera análisis de regresión lineal simple, destaca el gasto policial (Po1 y Po2) como una de las variables con mayor capacidad explicativa, lo que sugiere una dinámica reactiva de las instituciones:

Los estados con mayores tasas de criminalidad tienden a aumentar su inversión en recursos policiales.

En en análisis conjunto de las variables que hemos realizado, la relación entre desigualdad, nivel educativo e indicadores delictivos observada en nuestros modelos coincide con estudios posteriores que relacionan la estructura económica y los niveles de educación de una sociedad, con el comportamiento criminal. Investigaciones como la de Kelly (2000), por ejemplo, sugieren que la desigualdad de ingresos y la diferencia de nivele educativos están asociados con niveles más altos de delincuencia violenta y patrimonial.

Podemos concluir afirmando que los resultados coinciden en gran medida con la literatura clásica, pero también aportan matices relevantes sobre la fuerza relativa de los predictores cuando se analizan individualmente.

4.3.Implicaciones teóricas o prácticas.

Este estudio ofrece conclusiones valiosas tanto para la teoría como para la práctica. En el ámbito teórico, respalda la visión del delito como una conducta influenciada por la evaluación racional de costes y beneficios, propuesta inicialmente por Becker y ampliada por Ehrlich. La evidencia, especialmente en

variables relacionadas con la probabilidad de sanción o el nivel de recursos policiales, sugiere que la actividad delictiva no solo depende de factores sociales o psicológicos, sino también de cómo las instituciones configuran los incentivos y las posibles consecuencias si se opta por delinquir.

En el ámbito práctico, los resultados pueden guiar el diseño de políticas basadas en el Big Data. La relación significativa entre gasto policial y criminalidad destaca la importancia de una asignación estratégica de recursos, combinando acciones reactivas con medidas preventivas más amplias. La presencia de variables socioeconómicas significativas apunta a la necesidad de complementar las estrategias de seguridad con políticas centradas en aspectos como la educación, la cohesión social o la reducción de desigualdades, para abordar el fenómeno delictivo de manera más integral.

Por último, el estudio demuestra el potencial del Big Data y las técnicas de análisis de datos como herramientas útiles que pueden utilizarse en la gestión de la seguridad. La integración de grandes cantidades de datos con métodos de análisis avanzados permite comprender con con mucha más precisión las dinámicas delictivas, anticipar patrones y evaluar de forma más rigurosa la efectividad de las intervenciones públicas.

4.4.Limitaciones del estudio.

Si bien los resultados que hemos obtenido, sobre todo en el análisis cojuntos, son consistentes y ofrecen información valiosa sobre los factores asociados a la criminalidad, el estudio presenta varias limitaciones que es importante mencionar.

El tamaño de la muestra es relativamente pequeño (47 observaciones), lo que limita la generalización de los modelos y puede afectar a la estabilidad de los coeficientes que hemos estimado. Además, los datos se refieren a un periodo histórico específico (años 60), por lo que las conclusiones deben interpretarse en el contexto de esa época y no pueden extrapolarse directamente a situaciones actuales.

Por ultimo también debemos de mencionar dentro de las limitaciones del estudio, que solo se han usado técnicas de regresión lineales con lo que quedaría aplicar métodos más avanzados de análisis estadístico.

5. CONCLUSIONES

Este estudio ha tenido como objetivo principal determinar un modelo de regresión lineal capaz de predecir la tasa de criminalidad a partir de las variables económicas, sociales y policiales incluidas en el conjunto de datos analizado. Se ha seguido un proceso de análisis basado en la comparación de distintos modelos, comenzando por análisis simples y poco a poco se ha avanzado hacia propuestas más completas.

5.1.Resumen de los resultados más relevantes.

En primer lugar, se han aplicado modelos de regresión lineal simple para evaluar la relación lineal individual entre cada variable y la tasa de criminalidad. Estos primeros análisis nos han permitido ver el comportamiento de cada factor por separado. Hemos podido concluir que ninguno de los modelos simples ha resultado ser lo bastante significativo como para ofrecer una capacidad predictiva sólida. Aunque variables como el gasto policial (Po1 y Po2), el nivel educativo medio (Ed) y la probabilidad de reincidencia (Prob) han mostrado ciertos niveles de significación, su efecto aislado no ha permitido construir un modelo con un ajuste adecuado.

A continuación se ha elaborado un modelo de regresión lineal múltiple incorporando únicamente aquellas variables que, en los análisis dos a dos, han resultado más significantes. Este segundo modelo tampoco ha alcanzado un nivel satisfactorio de ajuste. Esto nos lleva a concluir que la selección manual y reducida de esas variables no representa de forma adecuada todos los factores que pueden integrar el fenómeno delictivo.

Con la construcción del tercer modelo, una regresión lineal múltiple completa en la que hemos integrado todas las variables del conjunto de datos,

hemos conseguido llegar al punto que queríamos. Este modelo sí nos ha ofrecido un ajuste adecuado. Posee una capacidad predictiva sólida y ha permitido identificar qué factores mantienen una influencia relevante cuando se analizan de manera conjunta.

Entre ellos ha destacado la variable de desigualdad (Ineq), cuya importancia refuerza la idea de que las desigualdades tanto económicas como sociales desempeñan un papel de gran relevancia en la explicación de la criminalidad. Este tercer modelo es, el modelo predictivo que se ha seleccionado, sobre el cual se han realizado las pruebas de predicción finales del trabajo al ser el más adecuado.

En cuanto a las predicciones efectuadas con el conjunto de prueba, se ha podido demostrar que su capacidad predictiva es adecuada, especialmente tratándose de datos reales y de un fenómeno complejo como la criminalidad.

Es cierto que los valores de MSE y RMSE han reflejado cierto error. No obstante, el modelo ha conseguido aproximarse a los valores reales y reproducir la tendencia general de la tasa de criminalidad como hemos podido observar además en el gráfico

Los resultados confirman que la criminalidad es un fenómeno en el que intervienen múltiples factores y cuya explicación no puede atribuirse a un solo elemento, sino a la interacción de componentes económicos, sociales y de gestión pública.

5.2.Evaluación del cumplimiento de los objetivos.

Los objetivos que nos planteamos al inicio del trabajo se han cumplido adecuadamente. Se ha logrado examinar el papel del Big Data en el análisis de fenómenos complejos, mostrando su potencial para integrar información de diversos tipos y obtener conocimiento útil a partir de grandes volúmenes de datos. Además, se ha destacado cómo estas capacidades resultan especialmente valiosas en ámbitos como la seguridad y la prevención del delito, donde el análisis

masivo de datos permite identificar patrones, anticipar riesgos y apoyar la toma de decisiones estratégicas en contextos relacionados con la criminalidad y la defensa
.

El uso del software R-Studio ha permitido aplicar modelos de regresión lineales simples y múltiples y evaluar la significancia estadística de cada variable, pudiendo ver qué factores del conjunto de datos podrían estar asociados a mayores tasas de criminalidad.

Por otro lado, se ha conseguido presentar un análisis detallado que facilita la interpretación de los resultados (utilizando R-markdonwn) y aporta una base para poder investigar posteriormente de forma más amplia o con modelos predictivos más complejos. Así pues, podemos decir que, el trabajo ha cumplido con el objetivo de ofrecer una aproximación empírica y fundamentada al estudio de la criminalidad desde una perspectiva basada en datos.

5.3.Propuestas futuras o recomendaciones.

Aunque este trabajo se ha centrado en un caso de estudio concreto, los métodos empleados pueden aplicarse a conjuntos de datos más grandes y más complejos, de hecho creo que mejoraría mucho los resultados obtenidos al aumentar el tamaño y la calidad de la muestra. Esto resulta especialmente relevante en contextos reales de aplicación a la seguridad y la defensa, donde el uso de Big Data se ha convertido en una herramienta esencial para procesar volúmenes masivos de información procedente de multitud de fuentes.

Para finalizar, podemos concluir afirmando que el trabajo pone de relieve el valor que aportan los enfoques basados en datos para interpretar mejor la información disponible y apoyar la toma de decisiones estratégicas, reforzando así el papel del análisis cuantitativo en la gestión moderna de la seguridad pública. La integración de Big Data y analítica avanzada se presenta, como un pilar fundamental para mejorar la eficacia, la anticipación y la capacidad operativa en el ámbito de la seguridad y la defensa.

6. REFERENCIA BIBLIOGRÁFICA

- Alteryx. (s.f.). Supervised vs. Unsupervised Learning: Which Is Best? Recuperado de https://www.alteryx.com/glossary/supervised-vs-unsupervised-learning
- Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Duxbury.
- De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. Library Review, 65(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061
- Dumbill, E. (January 11, 2012). What is big data? An introduction to the big data landscape. **O'Reilly Radar.** https://www.oreilly.com/radar/what-is-big-data/
- Ehrlich, I. (1973). Participation in illegitimate activities: A theoretical and empirical investigation. Journal of Political Economy, 81(3), 521–565.
- IBM. (2023). *Supervised vs. unsupervised learning: What's the difference?* IBM Think. https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning
- Kelly, M. (2000). Inequality and crime. Review of Economics and Statistics, 82(4), 530–539.
- Kröger, H., Stahl, A., & Unkelbach, J. (2022). *Regression: Linear Models in Statistics*. Springer.
- Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.
- McLeod, S. (August 11, 2025). Understanding P-Values and Statistical Significance. Simply Psychology. https://www.simplypsychology.org/p-value.html
- Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.
- Smith, J. (June 21, 2014). NYPD and Microsoft collaborate to create the Domain Awareness System (DAS). Medium. https://medium.com/homeland-security/nypd-and-microsoft-collaborate-to-create-the-domain-awareness-system-das-543a6245cb8f

IMÁGENES

- IMAGEN 1. OpenAI. (2025, octubre 21). *Las 5 V del Big Data* [Imagen generada por inteligencia artificial con ChatGPT (modelo GPT-5)]. OpenAI. https://chat.openai.com/
- IMAGEN 2. NYPD and Microsoft collaborate to create the Domain Awareness System (DAS). Fuente: https://medium.com/homeland-security/nypd-and-microsoft-collaborate-to-create-the-domain-awareness-system-das-543a6245cb8f