Procesamiento de Datos para el Análisis Social y Territorial de Nueva York

Entrega Final – Análisis y Recomendaciones

Autores

Julián Camilo Ramos Granada
María Fernanda Rodríguez Ospina
Sebastián Andrés Rodríguez Pérez

Profesor

John Corredor Franco, PhD

Pontificia Universidad Javeriana Facultad de Ingeniería
Departamento de Ingeniería de Sistemas
Curso: Procesamiento de Alto Volumen de Datos
Octubre de 2025

Índice

1. Introducción

El presente proyecto tiene como propósito aplicar técnicas de procesamiento de alto volumen de datos para analizar problemáticas sociales y urbanas en la ciudad de Nueva York. A partir de fuentes oficiales de datos abiertos, se busca comprender fenómenos asociados a la seguridad, la movilidad y las condiciones socioeconómicas de la población, con el fin de generar evidencia que contribuya a la toma de decisiones públicas. El trabajo se desarrolló empleando Apache Spark como herramienta principal para el procesamiento distribuido, la limpieza y el análisis de grandes volúmenes de información.

En la primera fase del proyecto se abordaron el entendimiento del negocio y de los datos, estableciendo el contexto urbano, seleccionando los conjuntos de información más relevantes y realizando la exploración inicial de los indicadores. Esta segunda entrega amplía el alcance del estudio al incorporar la preparación final de los datos, el modelamiento con técnicas de aprendizaje de máquina y la interpretación integrada de los resultados.

El proceso sigue la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), avanzando desde la preparación y modelado hasta la evaluación y generación de conclusiones. A partir del procesamiento realizado en el clúster Spark, se lograron identificar patrones de desigualdad territorial, relaciones entre pobreza, educación y criminalidad, así como factores asociados a la accidentalidad vial en la ciudad.

Finalmente, el documento presenta los hallazgos obtenidos y un conjunto de recomendaciones orientadas a mejorar la equidad urbana y la seguridad ciudadana. Con ello, se busca que el análisis realizado aporte una visión integral sobre las dinámicas sociales de Nueva York y sirva como base para futuras estrategias de planificación y desarrollo urbano sostenible.

2. Entendimiento del negocio

El entendimiento del negocio constituye la primera fase del proyecto y tiene como propósito contextualizar el caso de estudio en el entorno real de la ciudad de Nueva York. En esta sección se analiza la situación actual de la ciudad desde una perspectiva económica, social y territorial, identificando los principales retos que enfrenta en materia de seguridad, movilidad y equidad. Comprender este contexto es importante para orientar el análisis de datos hacia la generación de hallazgos útiles para la toma de decisiones públicas. Así, se parte del conocimiento del territorio y de sus indicadores clave para formular estrategias basadas en evidencia que contribuyan a mejorar los indicadores priorizados por el gobierno: la cantidad de arrestos y la frecuencia de accidentes viales.

2.1. Contexto general de Nueva York

Nueva York es una ciudad emblemática y una de las más influyentes a nivel mundial. Está situada en la costa noreste de Estados Unidos, en la desembocadura del río Hudson, frente al océano Atlántico. Su ubicación estratégica la convierte en un punto clave tanto para la economía estadounidense como para los negocios internacionales. La ciudad se destaca como el principal centro financiero del mundo, albergando la Bolsa de Valores de Nueva York y numerosas instituciones bancarias y financieras globales. Además, posee una riqueza cultural

y social extraordinaria gracias a la consentración de comunidades de todo el planeta, lo que ha moldeado su identidad diversa y multicultural. Esta mezcla cultural se refleja en sus museos, teatros, festivales y en la vibrante vida urbana que atrae a millones de turistas cada año. Nueva York también cumple un papel vital en la política global al ser sede de la Organización de las Naciones Unidas, reafirmando su importancia como un epicentro internacional para la cooperación y la diplomacia mundial.

La relevancia de Nueva York trasciende su economía, ya que representa un símbolo global de diversidad y resiliencia. La ciudad alberga sectores que abarcan desde las finanzas, la tecnología y el comercio, hasta las artes y los medios de comunicación, conformando un ecosistema urbano multifacético y dinámico. Su influencia se extiende a la cultura popular, la moda, la educación superior y la innovación tecnológica, lo que le otorga un lugar privilegiado en el escenario mundial. Conocida como "la ciudad que nunca duerme", Nueva York marca tendencias y movimientos sociales a nivel global, impactando las políticas económicas, las migraciones y la cultura contemporánea.

A pesar de su estatus y prosperidad, Nueva York enfrenta desafíos sociales significativos que afectan a diversos sectores de su población. La inseguridad, aunque ha disminuido considerablemente en las últimas décadas gracias a políticas públicas y estrategias de vigilancia, sigue siendo un reto en ciertas áreas urbanas donde los índices de criminalidad y violencia son más elevados, impactando especialmente a las comunidades más vulnerables. Asi mismo, los accidentes viales constituyen un problema social y de salud pública de gran importancia. A pesar de las medidas pioneras en seguridad vial y del liderazgo de la ciudad en la protección de peatones y ciclistas, las cifras de amenaza continúan generando preocupación, lo que exige esfuerzos constantes para mejorar la movilidad y la seguridad urbana.

Otro problema importante es la pobreza, visible en la marcada desigualdad económica entre distintos sectores de la ciudad. Mientras algunas zonas muestran altos niveles de vida y desarrollo, otras enfrentan limitaciones, falta de vivienda adecuada e inseguridad. Esta brecha socioeconómica repercute en el acceso y la calidad de la educación: a pesar de contar con una amplia oferta educativa y prestigiosas instituciones, las desigualdades afectan principalmente a los barrios más desfavorecidos.

Por lo tanto, Nueva York es una ciudad de contrastes donde conviven la innovación, la riqueza cultural y económica con desafíos sociales profundos. Su posición como nodo económico global y centro cultural de referencia es indiscutible, pero también lo es la importancia de abordar las problemáticas que afectan a su población para garantizar un desarrollo más equitativo y sostenible en el futuro.

2.2. Indicadores macroeconómicos de interés

Nueva York, con una población estimada en 8,48 millones de habitantes a julio de 2024, es una de las ciudades más densamente pobladas de Estados Unidos, lo que implica una gran concentración de recursos y desafíos urbanos. La población creció en aproximadamente 87.000 personas entre julio de 2023 y julio de 2024, un aumento que refleja una recuperación tras las pérdidas generadas por la pandemia, destacándose especialmente el crecimiento en los distritos de Manhattan y Brooklyn. Esta dinámica demográfica se ve influenciada tanto por la migración internacional como por movimientos internos dentro del país.

En materia de empleo, la tasa de desempleo en Nueva York para agosto de 2025 se

sitúa en torno al 4,7 %, ligeramente por encima del promedio nacional (4,3 %). Esto refleja una recuperación laboral gradual, aunque persisten brechas que afectan a ciertos sectores y grupos poblacionales, lo que demanda políticas activas de empleo y capacitación. La tasa de participación laboral, que mide la proporción de personas activas respecto a la población en edad de trabajar, se ubica en torno al 62,3 %.

En el ámbito educativo, la ciudad presenta un alto nivel de formación académica en su población. Para el año 2023, alrededor de 3,08 millones de residentes cuentan con un título universitario, mientras que 2,88 millones poseen educación secundaria completa. Asimismo, cerca de 1,77 millones han alcanzado un título de máster, 245.000 cuentan con un doctorado, y aproximadamente 373.000 personas no tienen educación formal. Estos datos reflejan una ciudad con una amplia base educativa, aunque las desigualdades sociales y económicas aún influyen en el acceso y la calidad de la formación en distintos barrios.

En cuanto a la pobreza, las Directrices Federales de Pobreza de Estados Unidos establecieron en 2023 un umbral de 14.580 dólares anuales para un hogar unipersonal y de 30.000 dólares para uno conformado por cuatro personas. Sin embargo, estos valores no se ajustan regionalmente, por lo que en una ciudad con un costo de vida elevado como Nueva York, las cifras oficiales pueden subestimar la magnitud real de la pobreza. Según estimaciones del Censo (ACS 2017–2021), aproximadamente el 17 % de la población —es decir, unos 1,5 millones de personas— vive por debajo del umbral federal de pobreza. De esta población, cerca del 29 % son menores de 18 años, y las mayores concentraciones de bajos ingresos se encuentran en el sur del Bronx, el Alto Manhattan y diversas zonas de Brooklyn.

En conjunto, estos indicadores evidencian una ciudad con un dinamismo económico y educativo destacado, pero también con desafíos persistentes en materia de desigualdad y acceso equitativo a oportunidades.

2.3. Objetivos

El propósito principal de este proyecto es apoyar al gobierno de la ciudad de Nueva York en la comprensión y mejora de dos indicadores prioritarios: la cantidad de arrestos y la frecuencia de accidentes viales. A partir del análisis de grandes volúmenes de datos públicos, se busca identificar patrones, factores asociados y posibles relaciones entre variables sociales, económicas y territoriales.

El objetivo es utilizar herramientas de procesamiento de datos a gran escala, específicamente Apache Spark, para generar hallazgos que sirvan como base para la toma de decisiones. Con esto se podrán diseñar soluciones concretas y eficientes que contribuyan a reducir los niveles de inseguridad y mejorar la movilidad en la ciudad.

Para esta primera entrega, el alcance del trabajo se centra en el entendimiento del negocio y de los datos disponibles. Esto incluye seleccionar las fuentes de información más relevantes, describir sus características, realizar una exploración inicial y formular preguntas de análisis que orienten las etapas posteriores del proyecto.

3. Selección de los datos a utilizar

Para el desarrollo del proyecto, se seleccionaron diversas fuentes oficiales de datos públicas del portal *Open Data NYC*, priorizando aquellas que permiten analizar los indicadores de interés definidos por el gobierno de Nueva York: la cantidad de arrestos y los accidentes viales. Adicionalmente, se incorporaron conjuntos de datos socioeconómicos y educativos que facilitan comprender las condiciones del entorno y su posible relación con los fenómenos en estudio.

3.1. Datasets propuestos por el gobierno

Los conjuntos de datos seleccionados y empleados durante esta etapa son los siguientes:

- **NYPD Arrest Data (Year-to-Date)**
  Fuente: `https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc`
  Contiene información sobre los arrestos realizados por el Departamento de Policía de Nueva York (NYPD), incluyendo tipo de delito, edad, sexo, raza, fecha y ubicación del suceso.

- **NYCgov Poverty Measure Data (2018)**
  Fuente: `https://data.cityofnewyork.us/City-Government/NYCgov-Poverty-Measure-Data-201cts7-vksw`
  Proporciona indicadores de pobreza, ingresos y distribución de la población por zonas geográficas.

- **Motor Vehicle Collisions – Vehicles**
  Fuente: `https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles-bm4k-52h4`
  Reúne información sobre vehículos involucrados en accidentes de tránsito, como tipo de vehículo, dirección del accidente, daños e información del conductor.

- **Health and Education Report 2016–2017**
  Fuente: `https://data.cityofnewyork.us/Education/2016-2017-Health-Education-Report/2dzy-e7cu`
  Incluye indicadores de bienestar y educación escolar en los distritos de Nueva York.

3.2. Justificación de selección de datos

Cada conjunto de datos aporta información específica para la resolución de los objetivos planteados. Los registros de arrestos permiten identificar patrones delictivos y zonas con mayor incidencia de detenciones, lo que permite crear estrategias de seguridad pública efectivas. Los datos de accidentes viales cuentan con evidencia cuantitativa sobre la siniestralidad en el tránsito urbano y sus causas facilitando la propuesta de medidas preventivas y de control. Por otro lado, la información sobre pobreza brinda el contexto económico y social necesario para analizar posibles correlaciones entre condiciones de vulnerabilidad y los indicadores de

seguridad o accidentes. Por último, los datos educativos y de salud escolar brindan una visión sobre los factores que pueden influir en la reducción de comportamientos delictivos y en la mejora de las condiciones sociales a largo plazo.

En general, la unión de estos datos permiten tener una visión integral del territorio y establecer relaciones entre variables sociales, económicas y de seguridad . Todo esto con el propósito de contar con información suficiente para formular soluciones de calidad , orientadas a reducir los niveles de inseguridad y la incidencia de accidentes en la ciudad de Nueva York.

4. Colección y descripción de datos

En esta etapa se llevó a cabo la recopilación, carga y revisión inicial de los conjuntos de datos seleccionados. Todos los procesos de lectura y análisis se realizaron sobre un clúster de procesamiento al cual se le realizó una configuración previa, utilizando Apache Spark como herramienta principal.

El objetivo de esta fase es garantizar que los datos estén correctamente integrados en el entorno de trabajo distribuido y que puedan ser procesados de manera eficiente. Para ello, se verificaron los formatos de los archivos, los tipos de datos y la estructura general de cada conjunto.

A continuación, se presentan las especificaciones del clúster utilizado y la descripción técnica de cada uno de los datasets empleados en el proyecto.

4.1. Características del clúster Spark

Para el procesamiento de los datos se implementó un clúster propio de Apache Spark sobre máquinas virtuales con sistema operativo Rocky Linux, configurado manualmente con el propósito de simular un entorno distribuido de cómputo y evaluar el rendimiento del procesamiento paralelo en tareas de análisis de datos a gran escala.

El clúster está conformado por tres nodos: un nodo *master* (que también actúa como *worker*) y dos nodos *worker* adicionales encargados de ejecutar las tareas distribuidas. Cada máquina virtual comparte la misma configuración de hardware, compuesta por 4 vCPU, 16 GB de memoria RAM y un disco de 80 GB configurado en modo *Thin Provision*, lo que permite reservar únicamente el espacio efectivamente utilizado. Todas las VMs operan bajo un entorno de red local privada definido dentro de la infraestructura de virtualización de la universidad.

En este esquema, la máquina *master* se encarga de la gestión y coordinación del clúster (despliegue de servicios, monitoreo y control), mientras que las dos *worker* aportan capacidad de cómputo adicional, recibiendo y ejecutando las tareas distribuidas. Esta configuración permite un balance adecuado entre realismo y eficiencia dentro de los recursos disponibles.

Desde el entorno de trabajo en Jupyter Notebook se estableció la sesión de Spark utilizando la biblioteca `PySpark`. La conexión se realizó mediante la inicialización del entorno con `findspark`, la importación de librerías principales y la creación de una sesión mediante `SparkConf()` y `SparkSession`. En esta configuración se definió el modo de planificación `FAIR`, el cual distribuye las tareas de forma equitativa entre los nodos disponibles.

4.2. Descripción técnica de los datasets

Una vez configurado el entorno de trabajo, se procedió a la carga y exploración inicial de los conjuntos de datos seleccionados. Cada dataset fue analizado a partir de su estructura, tipo de variables y volumen de registros con el fin de comprender su composición y evaluar la calidad de la información disponible.

A continuación, se describen las principales características técnicas de cada base de datos utilizada en el proyecto.

4.2.1. Datos de arrestos en Nueva York (NYPD Arrest Data)

El primer conjunto de datos corresponde al archivo `nypd_arrests.csv`, el cual contiene información detallada sobre los arrestos efectuados por el Departamento de Policía de Nueva York (NYPD). Los datos fueron obtenidos desde el portal oficial de datos abiertos de la ciudad de Nueva York y cargados en el clúster configurado para el proyecto.

El archivo contiene un total de **143.494 registros** distribuidos en **19 columnas**. Cada registro representa un arresto individual con información demográfica del implicado, el tipo de delito, la localización, entre otros. A continuación, se presenta la descripción general de sus variables principales:

- `ARREST_KEY`: identificador único del arresto.
- `ARREST_DATE`: fecha en que se efectuó el arresto.
- `PD_CD` / `KY_CD`: códigos de clasificación del delito según el NYPD.
- `PD_DESC` / `OFNS_DESC`: descripciones internas y estandarizadas del delito.
- `LAW_CODE` / `LAW_CAT_CD`: código legal violado y categoría del delito (F: felonía, M: delito menor, V: violación).
- `ARREST_BORO` / `ARREST_PRECINCT`: condado y precinto donde se realizó el arresto.
- `AGE_GROUP`, `PERP_SEX`, `PERP_RACE`: variables demográficas del arrestado.
- `Latitude`, `Longitude`: coordenadas geográficas del arresto.

Durante la revisión inicial se verificó que la mayoría de las columnas no presentan valores nulos. Las únicas excepciones fueron:

- `KY_CD`: 10 valores nulos.
- `LAW_CAT_CD`: 687 valores nulos (equivalentes al 0.48 % del total de registros).

Por otro lado, no se encontraron registros duplicados en el conjunto.

El dataset ofrece un alto nivel de detalle sobre los arrestos realizados en la ciudad, abarcando distintos tipos de delitos y características poblacionales. Mediante este conjunto de datos se busca analizar la distribución territorial de los incidentes y los perfiles más frecuentes de arrestos dentro de los diferentes distritos de Nueva York.

4.2.2. Datos de pobreza en Nueva York (NYCgov Poverty Measure Data)

El segundo conjunto de datos utilizado corresponde al archivo `nycgov_poverty_data.csv`, el cual contiene indicadores oficiales del nivel de pobreza en la ciudad de Nueva York. La información proviene del *NYC Center for Economic Opportunity* y se encuentra publicada en el portal de datos abiertos de la ciudad.

Cada registro representa un año de medición e incluye datos por grupo poblacional, tipo de hogar y zona geográfica. Los indicadores permiten analizar la evolución de la pobreza, la distribución del ingreso y el impacto de los costos de vida en los diferentes distritos de la ciudad.

Entre las variables más relevantes se incluyen:

- `Year`: año de referencia de la medición.
- `Poverty_Rate`: porcentaje de la población que vive bajo la línea oficial de pobreza.
- `Near_Poverty_Rate`: proporción de personas con ingresos apenas por encima del umbral de pobreza.
- `Deep_Poverty_Rate`: porcentaje de la población con ingresos extremadamente bajos.
- `Median_Income`: ingreso familiar medio ajustado por tamaño del hogar.
- `Borough`: distrito o zona geográfica correspondiente (Bronx, Brooklyn, Manhattan, Queens o Staten Island).
- `Population`: población total considerada para el cálculo.

El conjunto de datos cuenta con un total de 55 registros y 7 columnas. Este conjunto permite observar diferencias marcadas entre distritos. El Bronx, por ejemplo, presenta históricamente los índices más altos de pobreza, mientras que Manhattan y Staten Island registran los valores más bajos. Con esto, se busca abordar los temas de seguridad y movilidad desde una perspectiva socioeconómica, facilitando el análisis de posibles correlaciones entre pobreza, vulnerabilidad y niveles de criminalidad o accidentalidad en la ciudad.

4.2.3. Datos de accidentes viales (Motor Vehicle Collisions – Vehicles)

El tercer conjunto de datos corresponde al archivo `motor_vehicle_collisions_vehicles.csv`, el cual contiene información sobre los vehículos involucrados en accidentes de tránsito registrados en la ciudad. La fuente original es el portal de datos abiertos del NYPD, bajo la categoría de seguridad pública.

Cada registro representa un vehículo asociado a un evento de colisión, identificado por un código único. El conjunto de datos cuenta con más de 4 millones de registros y 25 columnas, siendo este uno de los más extensos y relevantes para el proyecto. La información se considera estructurada y cuenta con un nivel de detalle alto, abarcando variables técnicas, direccionales y descriptivas sobre el incidente.

Entre las variables principales se incluyen:

- COLLISION_ID: identificador único del evento de colisión.
- CRASH_DATE y CRASH_TIME: fecha y hora del accidente.
- VEHICLE_TYPE, VEHICLE_MAKE, VEHICLE_MODEL y VEHICLE_YEAR: características del vehículo involucrado.
- DRIVER_SEX, DRIVER_LICENSE_STATUS y TRAVEL_DIRECTION: información sobre el conductor y su comportamiento previo al accidente.
- POINT_OF_IMPACT, VEHICLE_DAMAGE y PUBLIC_PROPERTY_DAMAGE: detalles del impacto y nivel de daño.
- CONTRIBUTING_FACTOR_1 a CONTRIBUTING_FACTOR_5: factores que contribuyeron al accidente según los reportes del NYPD.

Durante la revisión inicial, se observó que el dataset incluye registros con valores nulos en algunas columnas descriptivas, principalmente en los factores contribuyentes (CONTRIBUTING_FACTOR_4 y CONTRIBUTING_FACTOR_5), e indica que no todos los reportes contienen información completa. No se identificaron duplicados en el campo COLLISION_ID.

Con este conjunto se busca identificar patrones de siniestralidad vial, tipos de vehículos más involucrados y condiciones asociadas a los accidentes.

4.2.4. Datos de educación y salud escolar (SAT NYC)

El último conjunto de datos corresponde al archivo SAT_Results_NYC.csv, el cual contiene los resultados promedio de las pruebas SAT aplicadas a estudiantes de educación secundaria en la ciudad de Nueva York durante el año 2012. La información fue descargada desde el portal *NYC Open Data*, cuya última actualización se registró el 26 de noviembre de 2024.

El archivo fue cargado exitosamente en el entorno de Spark, con un total de **478 registros** y **6 columnas**. A diferencia de los datasets anteriores, este conjunto tiene un tamaño menor, pero ofrece una visión agregada del desempeño académico de los estudiantes en tres áreas clave: lectura crítica, matemáticas y escritura. Esta información puede servir como referencia para explorar correlaciones entre el nivel educativo y las tasas de criminalidad o accidentalidad por distrito.

Las variables incluidas en el conjunto son las siguientes:

- DBN: identificador único de la escuela (District Borough Number).
- SCHOOL_NAME: nombre oficial de la institución educativa.
- Num_of_SAT_Test_Takers: número de estudiantes que presentaron el examen SAT.
- SAT_Critical_Reading_Avg_Score: puntaje promedio de lectura crítica.
- SAT_Math_Avg_Score: puntaje promedio de matemáticas.
- SAT_Writing_Avg_Score: puntaje promedio de escritura.

Durante la carga, se observó que las columnas correspondientes a los puntajes y al número de estudiantes fueron interpretadas como tipo `string`, lo cual indica la presencia de valores no numéricos. Este comportamiento se debe a la existencia de registros con caracteres como `'s'` o celdas vacías, que impiden la conversión automática a tipo numérico.

El análisis de calidad evidenció que no existen valores nulos explícitos (`null`) ni registros duplicados. Sin embargo, se identificaron **57 registros con valores no numéricos** en las columnas relacionadas con los puntajes y el número de participantes, específicamente en:

- `Num_of_SAT_Test_Takers`
- `SAT_Critical_Reading_Avg_Score`
- `SAT_Math_Avg_Score`
- `SAT_Writing_Avg_Score`

Estos valores serán tratados como datos faltantes en la fase de limpieza, mediante su conversión a `null`.

5. Exploración de los datos

La fase de exploración de los datos tiene como objetivo comprender en profundidad la estructura, el comportamiento y las características generales de cada conjunto de datos utilizado en el proyecto. A través de un análisis estadístico descriptivo y la generación de diversas visualizaciones, se buscó identificar patrones, tendencias y posibles anomalías que sirvan como base para el análisis posterior. Esta etapa permitió reconocer diferencias relevantes entre distritos, categorías demográficas y niveles socioeconómicos, así como evidenciar relaciones preliminares entre las variables de interés.

5.1. Análisis estadístico descriptivo

Con el propósito de comprender mejor la estructura y el comportamiento general de los datos, se realizó un análisis estadístico descriptivo sobre los cuatro conjuntos trabajados. Esta etapa busca identificar patrones relevantes, diferencias entre grupos y valores predominantes en las variables clave, sirviendo además como una verificación final de la calidad de los datos antes de avanzar hacia etapas más complejas como visualizaciones o modelado.

Arrestos en Nueva York (NYPD Arrest Data).
El análisis reveló que el grupo de edad con mayor incidencia de arrestos fue el de personas entre 25 y 44 años, seguido por los adultos de 45 a 64. En contraste, los menores de edad y los adultos mayores de 65 años presentaron proporciones significativamente menores. Al examinar la variable de sexo, se observó que la diferencia es bastante evidente: más del 80 % de los arrestos fueron a hombres, mientras que las mujeres representaron poco menos del 20 %. Por su parte, la variable de raza mostró que las personas identificadas como negras y como hispanas (tanto blancas como negras) concentraban la mayoría de los registros. En términos territoriales, Brooklyn, Manhattan y el Bronx fueron los distritos que agruparon el

mayor volumen de arrestos, siendo los precintos 14, 40 y 75 algunos de los más recurrentes. Finalmente, al observar la variable de tipo de delito, se encontró que las agresiones menores, el hurto y los delitos relacionados con drogas fueron los más comunes. Cerca del 60 % de los delitos correspondían a la categoría legal de "delito menor" (M), mientras que el 40 % restante fueron clasificados como felonías (F).

Resultados académicos (SAT NYC).
Este conjunto contiene los puntajes promedio del examen SAT para 421 instituciones educativas de la ciudad. Una vez procesado el conjunto y corregidos los errores de tipado, se realizó un análisis de los puntajes por área. Los resultados generales se resumen en la siguiente tabla:

Área	**Media**	**Mediana**	**Máximo**
Lectura crítica	400.9	391	679
Matemáticas	413.4	395	735
Escritura	394.0	381	682

Como se puede evidenciar, las tres áreas mostraron un rendimiento similar, aunque matemáticas tuvo un desempeño ligeramente más alto. Las distribuciones fueron relativamente simétricas, con medianas cercanas a las medias. Sin embargo, también se identificaron escuelas con puntajes muy por debajo o muy por encima del promedio, lo que refleja diferencias significativas entre instituciones.

El número de estudiantes que presentó el examen varió ampliamente. Si bien el promedio fue de 110 estudiantes por colegio, la mediana fue de solo 62, lo que indica una distribución asimétrica. Hubo instituciones con menos de 10 participantes y otras con más de 1.200, lo que puede estar asociado al tamaño de la institución.

Accidentes viales (Motor Vehicle Collisions).
Este conjunto es el más extenso, con más de 4 millones de registros. Luego de su limpieza, se trabajó sobre 17 variables seleccionadas. Entre los tipos de vehículos más involucrados se encuentran los sedanes, station wagons, SUVs y taxis. Cabe destacar que una parte considerable de los registros originalmente no especificaba el tipo de vehículo, con etiquetas como *unknown* o *unspecified*, las cuales fueron unificadas bajo la categoría **NO_INFO** para facilitar su análisis y contabilización. Esta consolidación permitió visualizar con mayor claridad la proporción de datos faltantes en variables clave sin dispersar los resultados en múltiples formas de "sin información".

En cuanto a las causas de los accidentes, más del 60 % de los registros indicaban que el "factor contribuyente" era "no especificado". Estos también fueron agrupados bajo la misma etiqueta **NO_INFO**. Aun así, dentro de los factores que sí fueron reportados, los más frecuentes fueron la distracción del conductor, el no ceder el paso, y el exceso de velocidad. Estos resultados sugieren que una buena parte de los accidentes podría estar relacionada con comportamientos evitables.

También se analizaron los años de fabricación de los vehículos. El promedio fue cercano a 2014, con una mediana similar. Sin embargo, se detectaron valores extremos como 1000 o 20063, lo cual claramente corresponde a errores de ingreso. En cuanto a la ubicación del impacto, los golpes frontales fueron los más comunes, seguidos por impactos laterales y traseros. De forma similar a lo anterior, los valores no informados en esta categoría fueron

homogeneizados como **NO_INFO**. Solo un pequeño porcentaje de los registros confirmó daños a propiedad pública, mientras que la gran mayoría marcó esta variable como "no" o "desconocido", también agrupados bajo la misma etiqueta para mantener consistencia en el tratamiento de los datos faltantes.

Condiciones de pobreza (Microdatos NYCgov).
El conjunto de pobreza utilizado corresponde a una base de microdatos individuales. En total se analizaron 68.273 registros, distribuidos principalmente entre los distritos de Brooklyn, Queens y el Bronx. Al usar la variable `NYCgov_Pov_Stat`, que identifica si una persona se encuentra bajo la línea de pobreza, se estimó una tasa de pobreza general del 17.7 %. La distribución por distrito se resume en la siguiente tabla:

Distrito	Código	Tasa de pobreza (%)
Bronx	1	25.8
Brooklyn	2	18.3
Manhattan	3	13.8
Queens	4	15.8
Staten Island	5	13.8

El Bronx se consolida como el distrito con mayor proporción de personas en situación de pobreza, mientras que Manhattan y Staten Island presentan las tasas más bajas.

En cuanto al ingreso ajustado por la metodología del gobierno de la ciudad (`NYCgov_Income`), el promedio fue de aproximadamente $79.120, pero la mediana fue notablemente más baja ($60.122), lo que indica una distribución desigual. Se observaron valores extremos, desde ingresos negativos hasta cifras que superaban los $899.000. Estos casos podrían corresponder a familias con múltiples fuentes de ingreso.

En conjunto, este análisis permitió conocer mejor la forma en que están estructurados los datos y los patrones generales presentes en cada conjunto. Las diferencias encontradas entre distritos, categorías demográficas y niveles de ingreso generan una base para continuar con la fase de visualización e interpretación de correlaciones entre variables sociales, educativas y territoriales.

5.2. Visualizaciones

A partir de los distintos conjuntos de datos utilizados, se generaron gráficos que complementan los análisis estadísticos previos y facilitan la comprensión de fenómenos sociales, territoriales y demográficos en la ciudad.

En cada caso se seleccionaron variables clave, tanto cuantitativas como categóricas, y se representaron mediante histogramas, gráficos de barras, mapas de calor o boxplots, dependiendo del tipo de dato y del objetivo de análisis.

Cada subsección presenta una serie de cinco visualizaciones por conjunto de datos (arrestos, accidentes, SAT, pobreza), acompañadas de una breve interpretación que resalta los principales hallazgos. Estas gráficas permiten comparar distritos, identificar rangos etarios más afectados, explorar distribuciones económicas, evaluar diferencias raciales, y detectar comportamientos atípicos o categorías dominantes.

5.3. NYPD Arrest Data

Las visualizaciones seleccionadas para este conjunto buscan resaltar los aspectos más representativos de los arrestos en la ciudad durante 2025, complementando el análisis estadístico previo. Se eligieron variables clave como edad, sexo, tipo de delito, distrito y raza del arrestado, ya que permiten identificar patrones demográficos y geográficos relevantes. Además, se incluyó una gráfica que diferencia entre delitos mayores y menores, para tener una visión más clara del tipo de conductas sancionadas.

5.3.1. Distribución de arrestos por grupo de edad y sexo

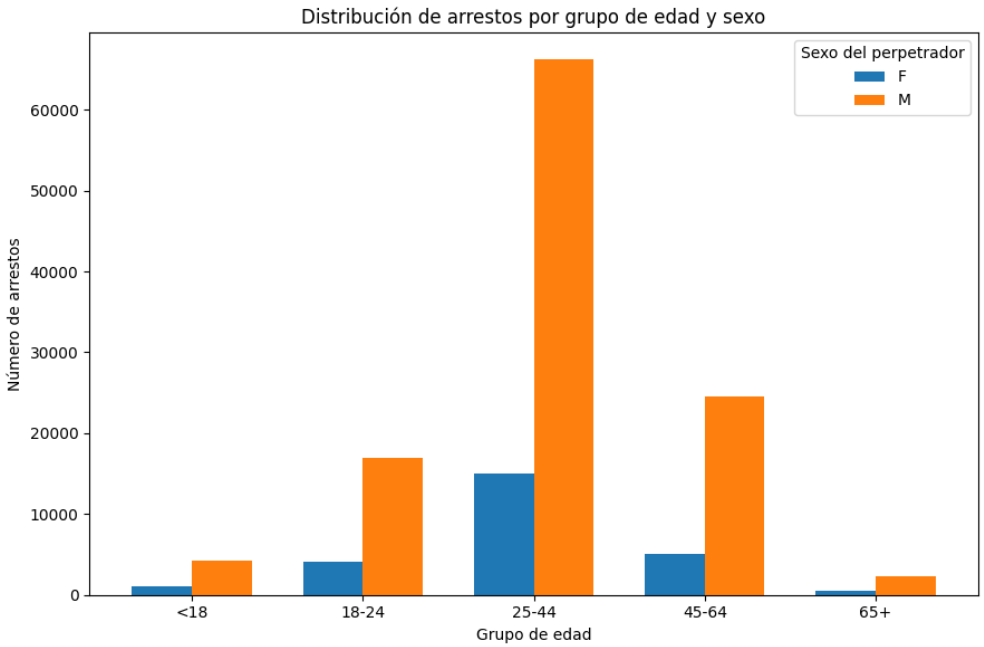

Figura 1: Distribución de arrestos por grupo de edad y sexo.

Este gráfico muestra la cantidad de arrestos según distintos grupos etarios, diferenciando por sexo. Se observa que la mayoría de los arrestos ocurre entre personas de 25 a 44 años, con una notable predominancia de hombres en todos los rangos de edad.

5.3.2. Top 15 delitos más frecuentes

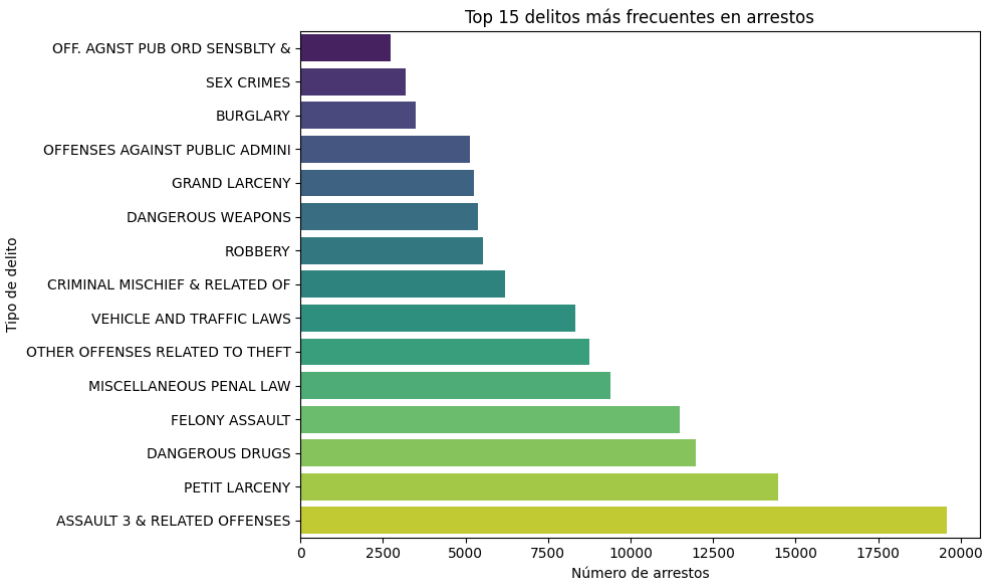

Figura 2: Quince delitos más frecuentes en los arrestos.

Se representa la frecuencia de arrestos según el tipo de delito. Encabezan la lista el asalto en tercer grado, el hurto menor y los delitos relacionados con drogas.

5.3.3. Número de arrestos por distrito (borough)

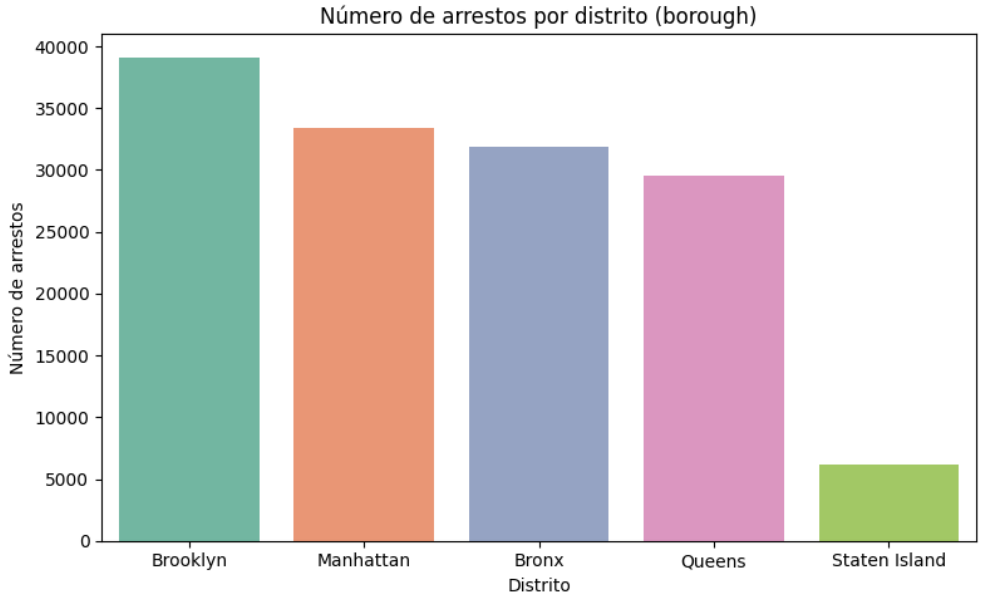

Figura 3: Total de arrestos por borough.

Brooklyn es el distrito con mayor cantidad de arrestos, seguido por Manhattan, Bronx y Queens. Esta distribución territorial puede estar asociada a factores como densidad poblacional o características socioeconómicas.

5.3.4. Distribución de arrestos por raza

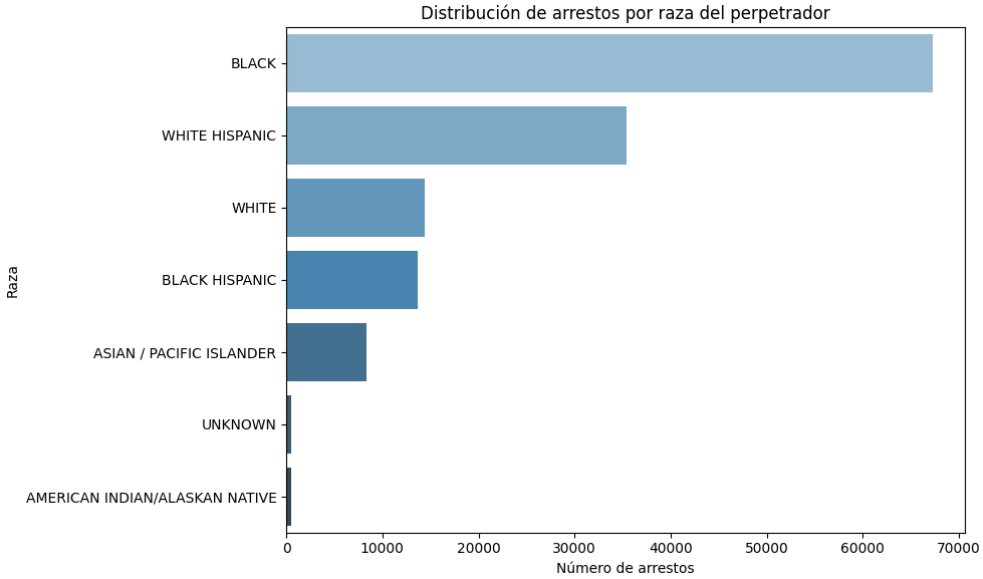

Figura 4: Distribución de arrestos según la raza del perpetrador.

En esta visualización se aprecia que las personas identificadas como Black y White Hispanic son las más representadas en los registros de arresto.

5.3.5. Distribución por tipo legal del arresto

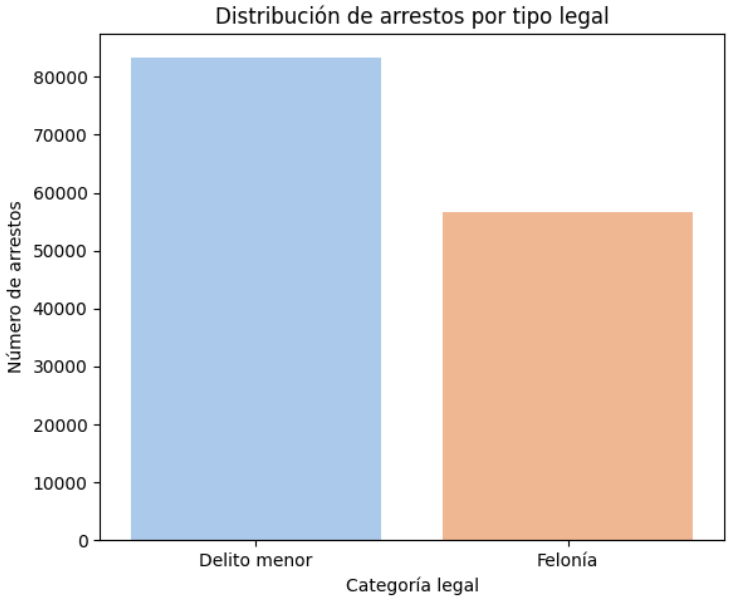

Figura 5: Distribución por tipo legal: felonía o delito menor.

La mayoría de los arrestos registrados corresponden a delitos menores. Esta gráfica ayuda a dimensionar qué tan grave es, en promedio, la conducta sancionada, y puede ser útil para reflexionar sobre prioridades en políticas de justicia y seguridad.

5.4. Motor Vehicle Collisions

En el caso del conjunto de accidentes vehiculares registrados en la ciudad de Nueva York, las visualizaciones permiten observar con mayor detalle las características más relevantes de los siniestros y las tendencias que se desprenden de ellos. Estas gráficas complementan los resultados estadísticos al representar de forma visual los tipos de vehículos más involucrados, las causas más comunes, las zonas de impacto, la condición de los vehículos al momento del choque y la evolución temporal de los accidentes.

Cabe señalar que, durante el proceso de limpieza, se identificó una gran cantidad de registros con valores no informados o ambiguos en distintas columnas. Por coherencia y para facilitar la interpretación, todas las categorías de este tipo fueron unificadas bajo la etiqueta `NO_INFO`. Esto permitió manejar de manera más clara los datos faltantes y distinguir entre la información efectivamente registrada y aquella que no fue reportada.

5.4.1. Tipos de vehículos más involucrados

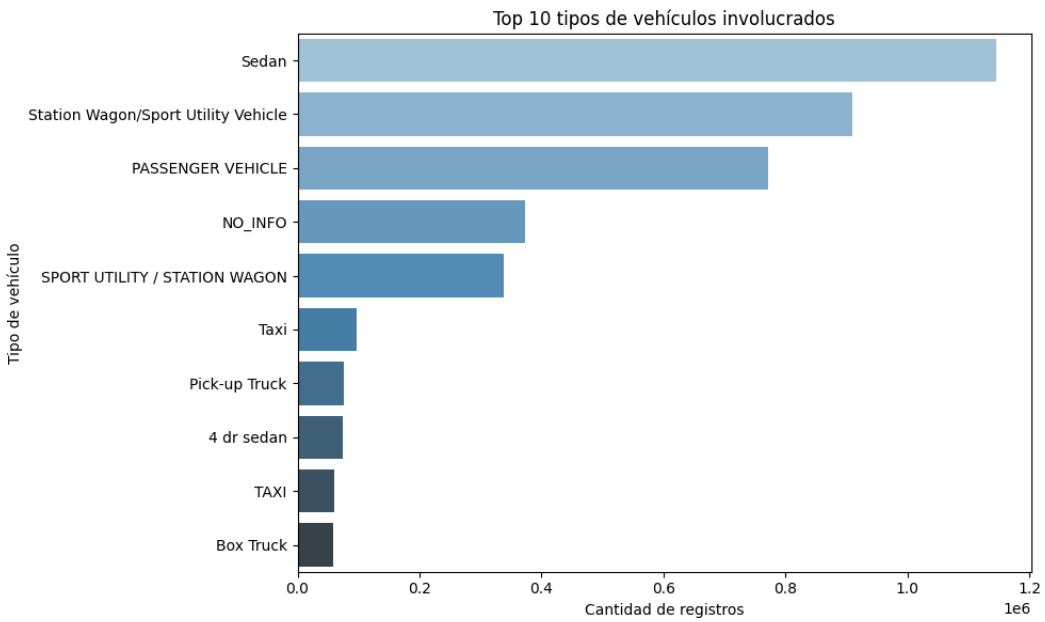

Figura 6: Top 10 tipos de vehículos más involucrados en accidentes.

El primer gráfico muestra los tipos de vehículos más frecuentemente implicados en colisiones. Los automóviles tipo *sedán* encabezan la lista, seguidos por los *station wagons* o SUVs y los vehículos clasificados genéricamente como *passenger vehicle*. La presencia de un número considerable de casos marcados como `NO_INFO` refleja limitaciones en la calidad del registro y la falta de precisión en algunos reportes, aunque no altera las tendencias generales. Este resultado indica que los vehículos de uso cotidiano son los que más se ven expuestos al riesgo de colisión dentro del entorno urbano.

5.4.2. Principales factores contribuyentes

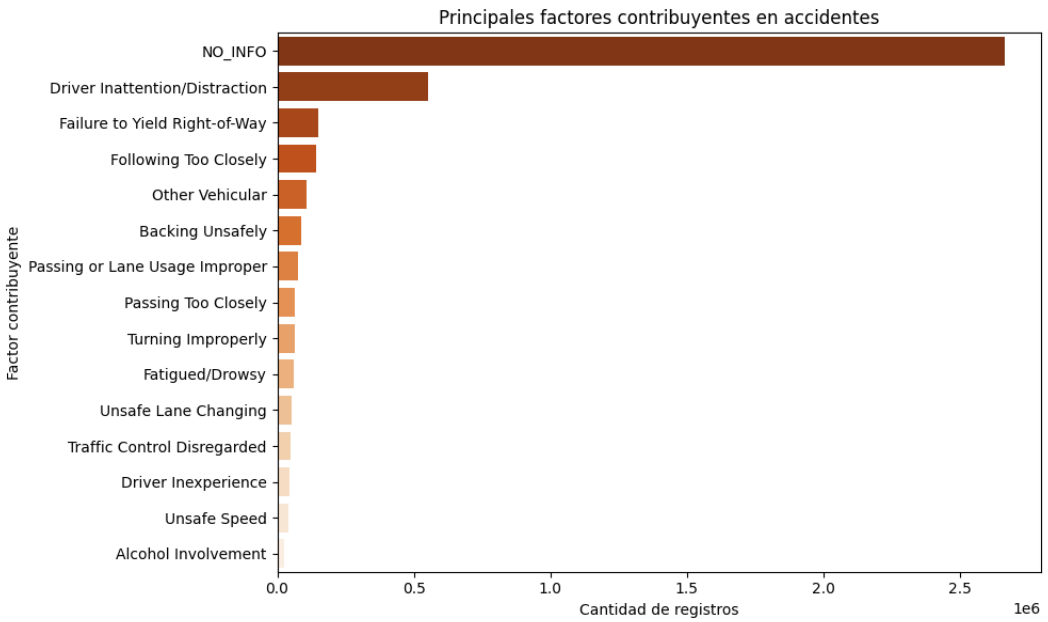

Figura 7: Factores contribuyentes más frecuentes en los accidentes.

El segundo gráfico ilustra los factores atribuidos como causa principal de los accidentes. Aunque la categoría `NO_INFO` concentra una proporción significativa de registros, entre los casos reportados sobresalen la distracción del conductor, el no ceder el paso y la conducción a corta distancia respecto a otros vehículos. Estas causas evidencian que la mayoría de los accidentes tienen un origen humano, asociado a fallas de atención, imprudencia o desconocimiento de las normas básicas de seguridad vial.

5.4.3. Puntos de impacto más comunes

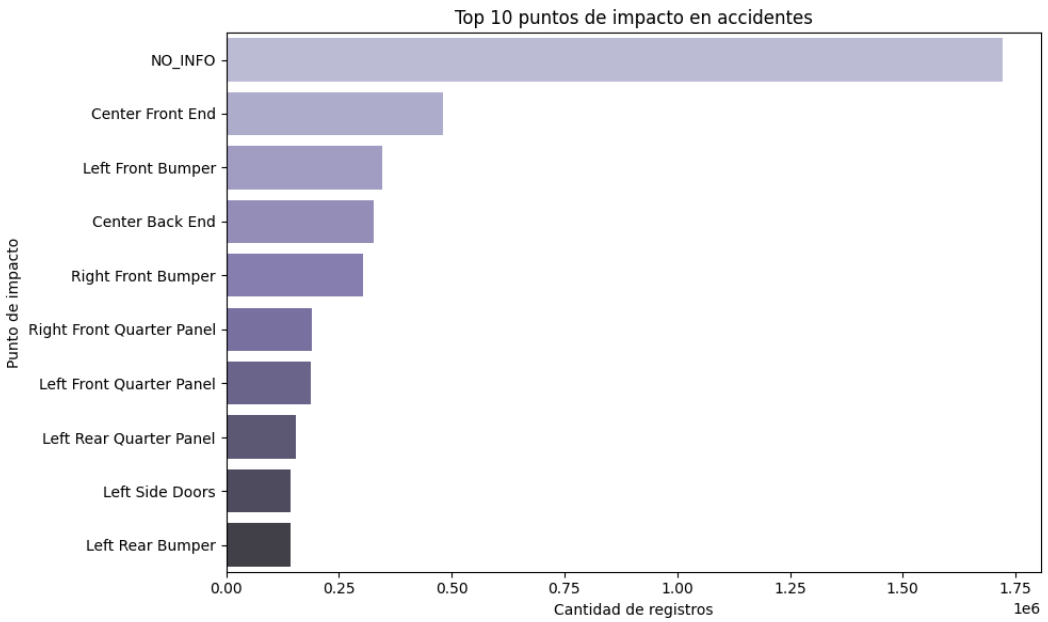

Figura 8: Distribución de puntos de impacto en los vehículos involucrados.

En este gráfico se observan las zonas del vehículo más afectadas por los choques. Los impactos frontales son los más frecuentes, seguidos por los golpes en las partes traseras y laterales. Este patrón coincide con la naturaleza de los siniestros más comunes, como colisiones por alcance o choques frontales en intersecciones. La abundancia de valores `NO_INFO` también pone en evidencia la dificultad de registrar de manera precisa ciertos detalles técnicos en los informes policiales.

5.4.4. Condición del vehículo antes del choque

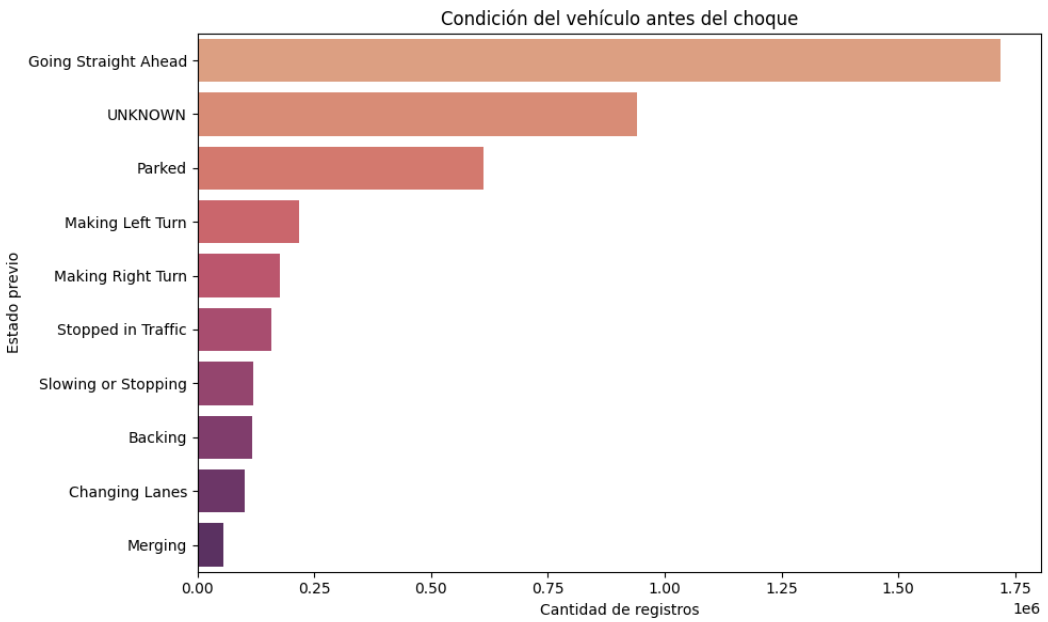

Figura 9: Condición del vehículo antes del choque.

La mayoría de los vehículos involucrados en accidentes estaban circulando en línea recta al momento del impacto, seguidos por aquellos que se encontraban estacionados o realizando maniobras de giro. Esto indica que muchos incidentes se producen en condiciones normales de desplazamiento, posiblemente por distracciones o fallas en la atención del conductor.

5.4.5. Evolución de colisiones por año

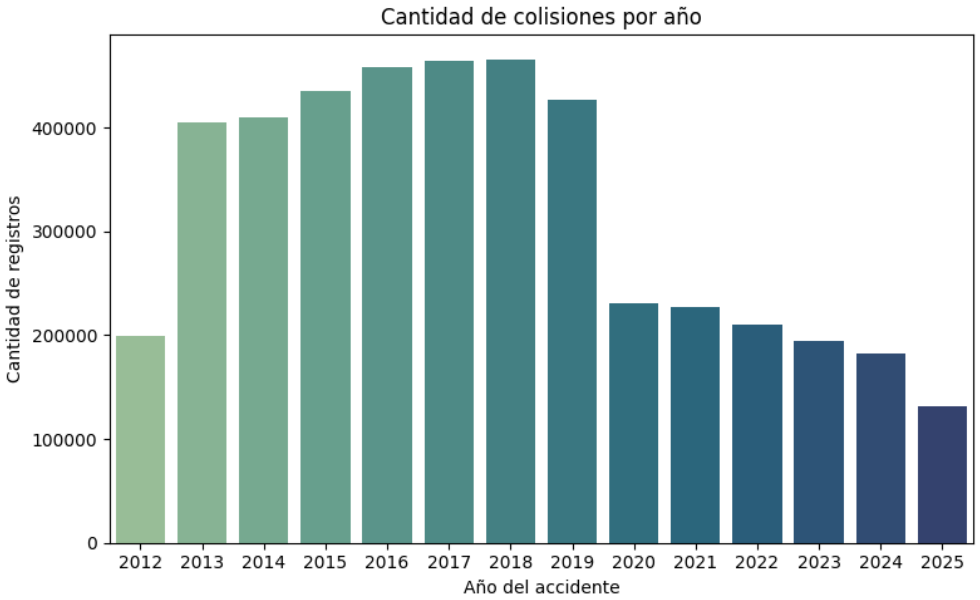

Figura 10: Evolución del número de colisiones registradas por año.

Finalmente, la última gráfica muestra la evolución anual del número de colisiones registradas. Se observa un incremento sostenido entre 2013 y 2019, seguido por una disminución marcada a partir de 2020. Este descenso podría estar asociado a la pandemia de COVID-19 y las restricciones de movilidad que redujeron la circulación de vehículos,. En conjunto, esta visualización proporciona una perspectiva temporal que ayuda a contextualizar las variaciones en la frecuencia de accidentes dentro del periodo analizado.

5.5. Poverty Data

A partir de esta base de datos, se construyeron diversas visualizaciones que permiten observar con mayor claridad la distribución de la pobreza en la ciudad, así como su relación con variables como el ingreso familiar, el distrito de residencia y el nivel educativo.

5.5.1. Cantidad de registros por distrito

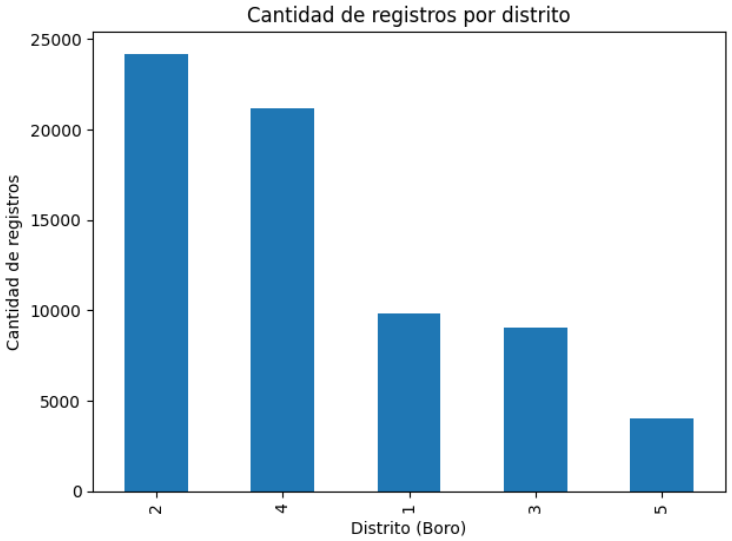

Figura 11: Cantidad de registros por distrito.

Este gráfico muestra el volumen de registros en cada uno de los cinco distritos (boroughs) de Nueva York. Aunque no representa una variable analítica directamente, es útil para tener contexto sobre la cobertura y densidad de información proveniente de cada zona. Los distritos 2 y 4 concentran el mayor número de registros, seguidos por el 1, 3 y 5.

5.5.2. Tasa de pobreza por distrito

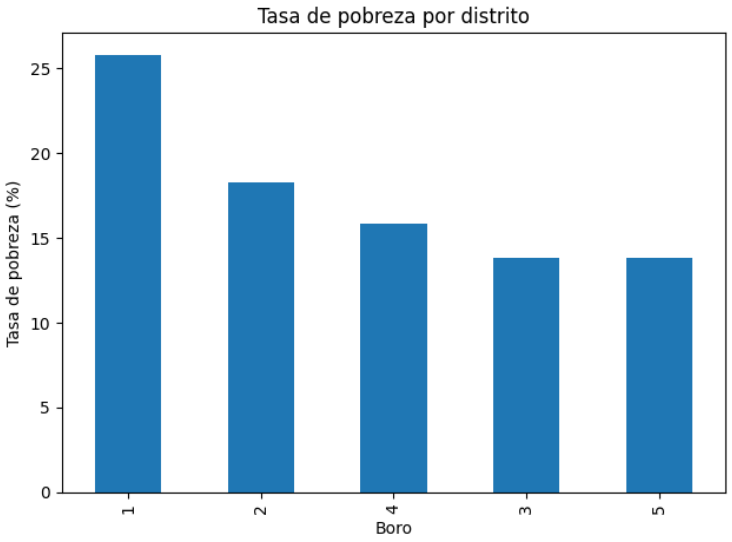

Figura 12: Tasa de pobreza por distrito.

Este gráfico presenta el porcentaje de personas clasificadas como en condición de pobreza, según su distrito de residencia. Se evidencia una desigualdad notable, siendo el distrito 1 el que presenta la mayor tasa de pobreza (por encima del 25 %), mientras que los distritos 3 y 5 muestran los niveles más bajos. Esta información es clave para identificar zonas con mayores necesidades económicas y orientar estrategias de intervención.

5.5.3. Distribución de ingresos familiares estimados

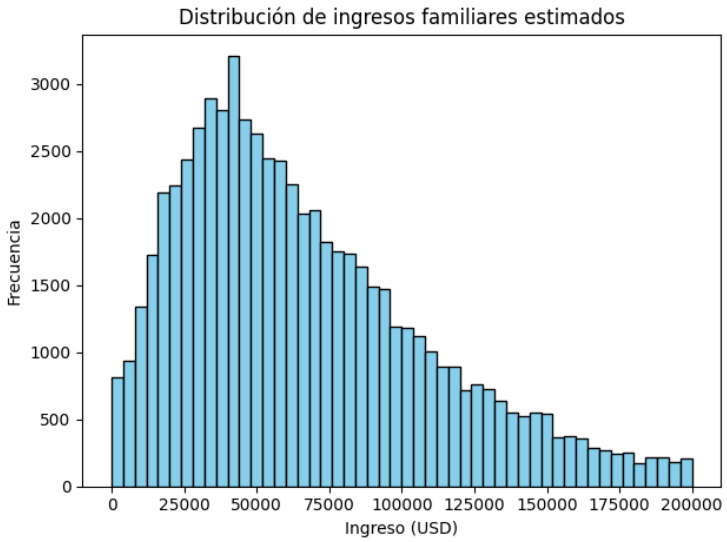

Figura 13: Distribución estimada de ingresos familiares.

El histograma de ingresos familiares muestra que la mayoría de los hogares se concentran en rangos bajos o medios, particularmente entre los $20.000 y $60.000 anuales. A medida que el ingreso aumenta, la frecuencia disminuye, reflejando una distribución desigual donde los ingresos elevados son mucho menos comunes.

5.5.4. Boxplot de ingreso familiar

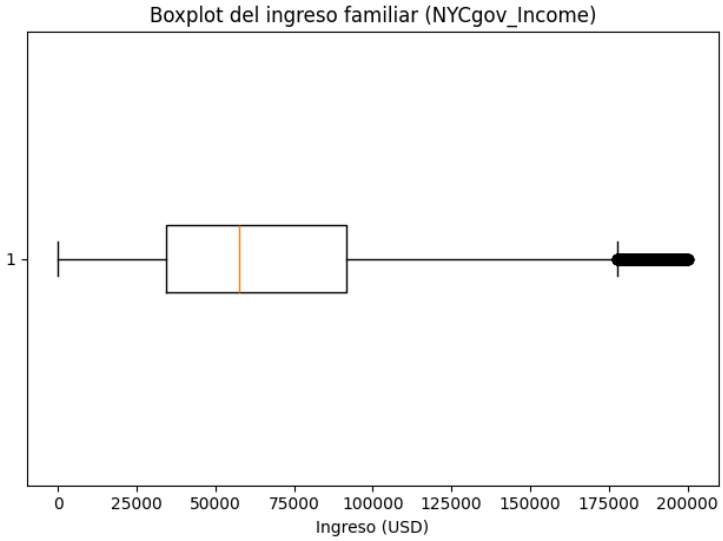

Figura 14: Boxplot del ingreso familiar estimado.

El boxplot refuerza lo observado anteriormente, mostrando una mediana cercana a los $55.000 y una amplia dispersión de los datos. También se evidencian múltiples valores atípicos hacia el extremo superior, lo cual sugiere la presencia de hogares con ingresos significativamente más altos que el promedio general.

5.5.5. Ingreso por nivel educativo

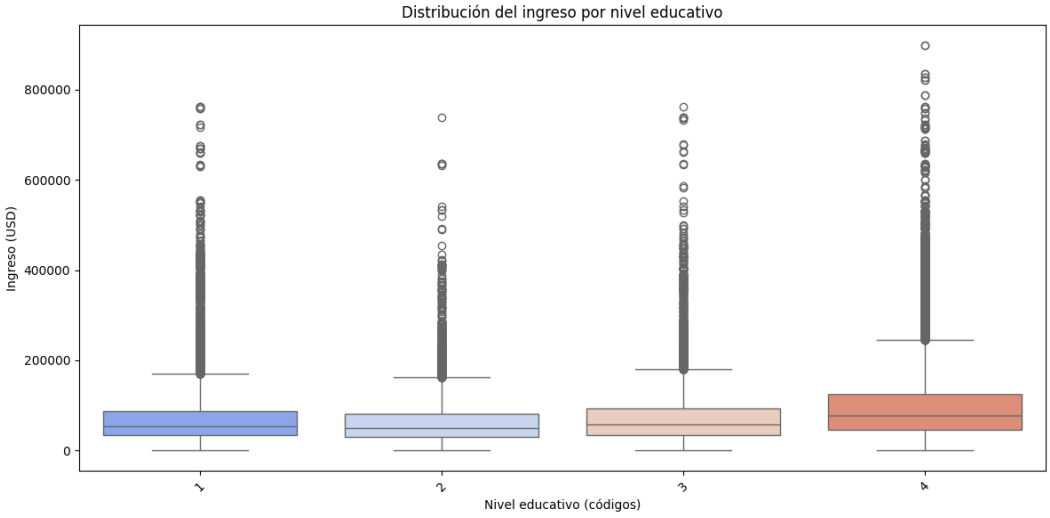

Figura 15: Distribución del ingreso familiar según nivel educativo.

Este gráfico compara el ingreso estimado entre los distintos niveles educativos registrados. A mayor nivel educativo, mayor es la mediana de ingreso y mayor es también el rango intercuartílico. Esto sugiere una clara relación entre el nivel educativo y el bienestar económico.

5.6. SAT NYC

En el caso del conjunto de datos de resultados del SAT para escuelas públicas de Nueva York, se elaboraron visualizaciones que permiten examinar con mayor detalle el rendimiento promedio de los estudiantes en las tres áreas evaluadas: lectura crítica, matemáticas y escritura. Estas gráficas complementan el análisis estadístico previo, facilitando la identificación de patrones y posibles correlaciones entre los puntajes.

Se incluyen histogramas para cada componente del examen, un boxplot comparativo y un gráfico de dispersión que permite observar la relación entre lectura y matemáticas.

5.6.1. Distribución de puntajes de lectura crítica

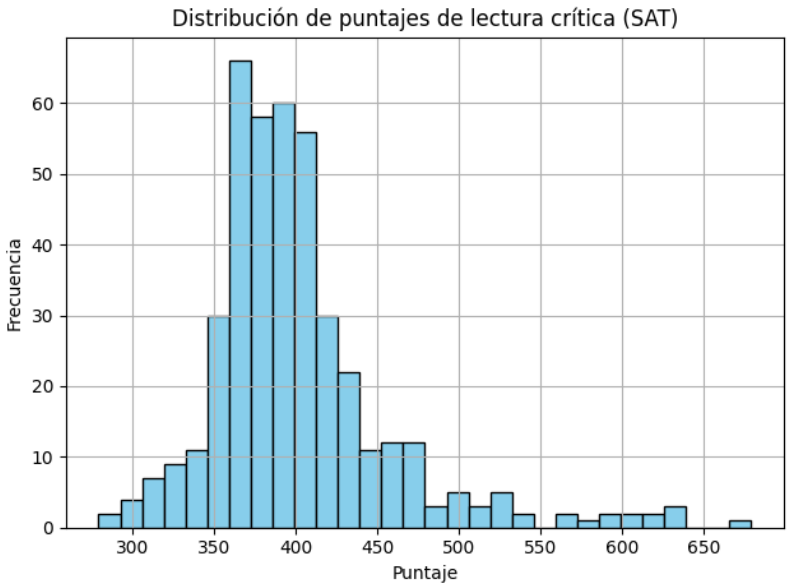

Figura 16: Distribución de puntajes de lectura crítica (SAT).

El histograma refleja que la mayoría de las escuelas presentan puntajes promedio entre 350 y 400 puntos en lectura crítica. Se trata de una distribución ligeramente sesgada hacia la derecha, con pocas instituciones alcanzando valores cercanos o superiores a 600 puntos.

5.6.2. Distribución de puntajes de matemáticas

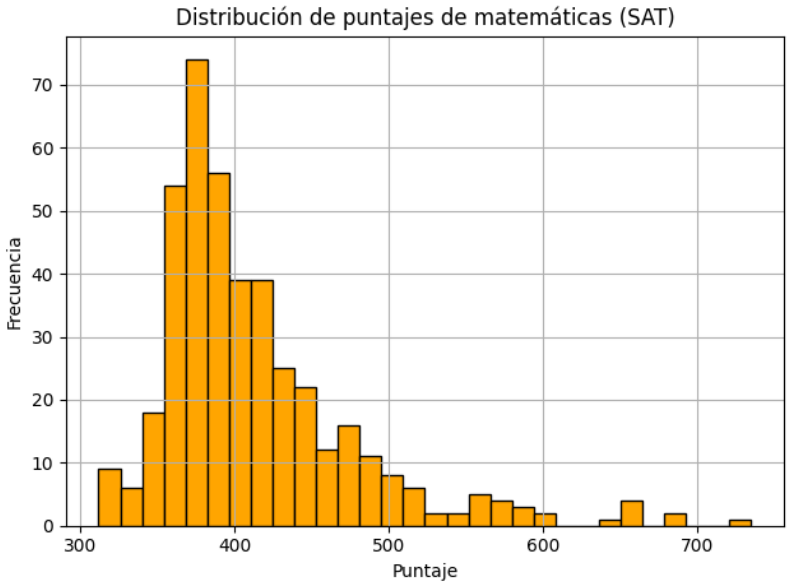

Figura 17: Distribución de puntajes de matemáticas (SAT).

La distribución en matemáticas muestra una mayor dispersión, con una concentración similar a la de lectura en el rango medio, pero con más casos que alcanzan puntajes elevados, superando incluso los 700 puntos. Esto indica un mejor desempeño en esta sección.

5.6.3. Distribución de puntajes de escritura

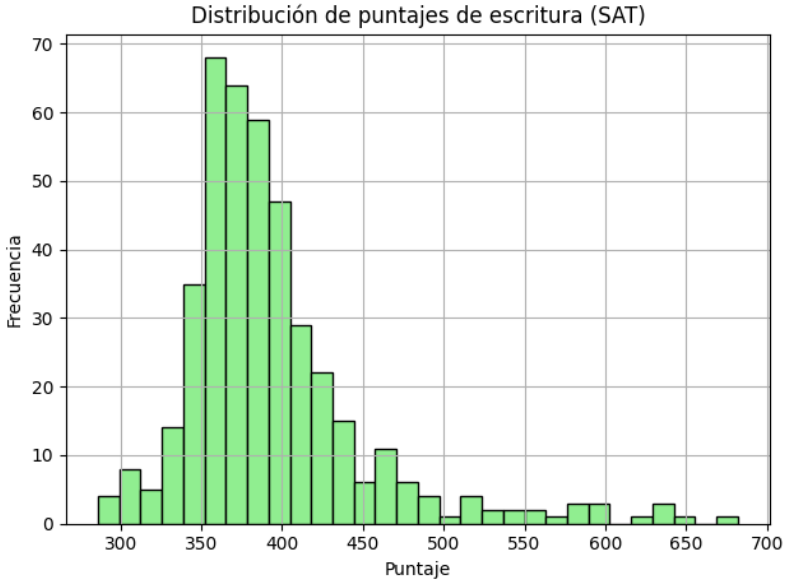

Figura 18: Distribución de puntajes de escritura (SAT).

Los resultados de escritura presentan una forma similar a la de lectura crítica, con una fuerte concentración entre los 350 y 400 puntos. La menor dispersión sugiere una menor variabilidad en el desempeño de las escuelas en esta área.

5.6.4. Comparación de puntajes por componente

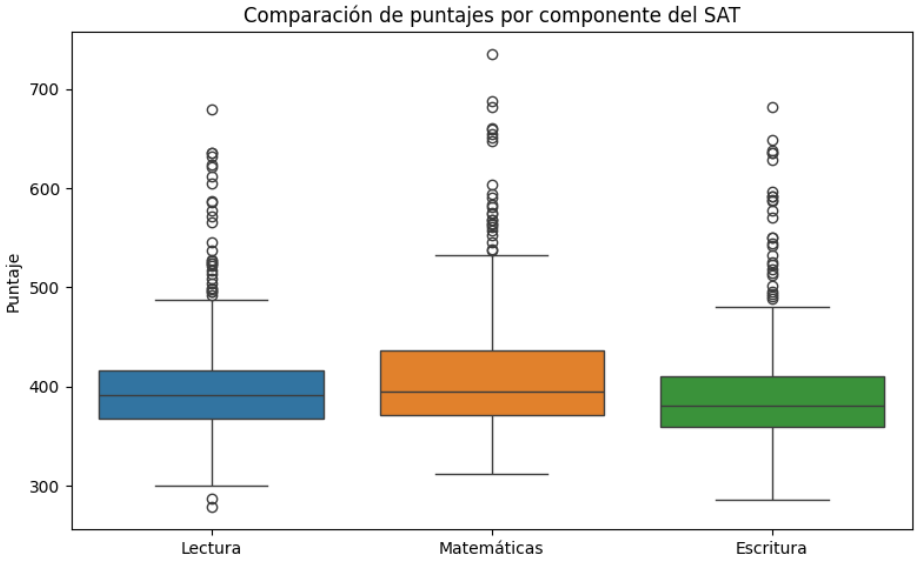

Figura 19: Comparación de puntajes por componente del SAT.

Este boxplot permite comparar visualmente la distribución de puntajes entre las tres secciones del examen. Se observa que matemáticas presenta una mediana ligeramente superior, así como una mayor presencia de valores atípicos.

5.6.5. Relación entre puntajes de lectura y matemáticas

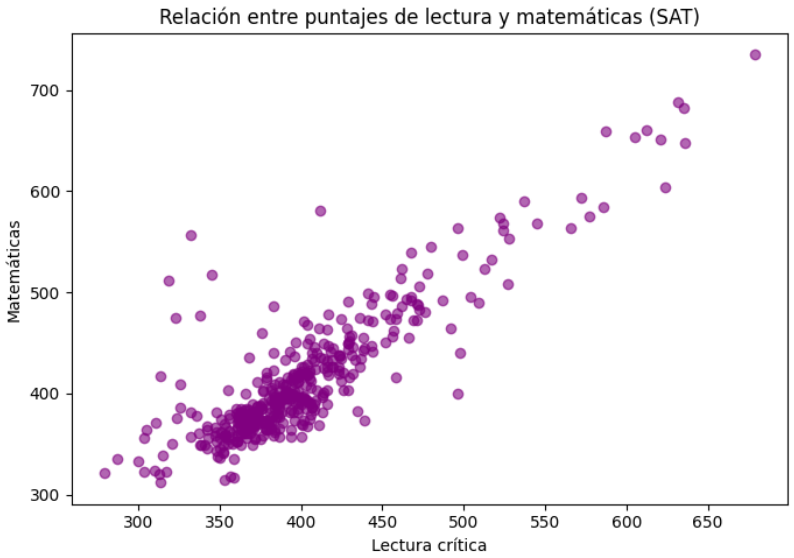

Figura 20: Relación entre puntajes de lectura crítica y matemáticas.

El gráfico de dispersión revela una correlación positiva entre los puntajes en lectura y matemáticas. Las escuelas con buen desempeño en una de estas áreas tienden también a obtener buenos resultados en la otra, lo que sugiere la posible existencia de factores comunes que afectan ambas dimensiones del rendimiento académico.

5.7. Hallazgos preliminares

Durante la fase de exploración de los datos, se identificaron una serie de patrones que pueden aportar al análisis territorial y social de la ciudad. Si bien aún no se ha realizado un cruce formal entre los conjuntos, algunos hallazgos que destacan son:

- **Concentración etaria en los arrestos:** La mayoría de los arrestos se concentran en personas entre los 25 y 44 años, con una proporción mucho mayor de hombres que de mujeres.
- **Alta presencia de registros no informados en accidentes:** En el conjunto de accidentes vehiculares, varias columnas presentan valores desconocidos o sin especificar. Aun así, se evidencian causas frecuentes como distracción del conductor o maniobras inseguras. También se detectó una alta proporción de impactos frontales y vehículos sedán como los más comúnmente involucrados.
- **Diferencias distritales en pobreza:** En el análisis del conjunto de pobreza, se observan diferencias importantes entre los distritos. Por ejemplo, Manhattan (distrito 1) presenta la tasa de pobreza más alta, mientras que Staten Island (distrito 5) tiene la más baja. Además, se aprecia que las personas en condición de pobreza tienen ingresos medios sustancialmente menores, como era de esperarse.

- **Relaciones entre desempeño académico:** El conjunto SAT muestra una correlación clara entre los puntajes en lectura crítica y matemáticas. Las escuelas que tienen mejores resultados en una sección tienden a tenerlos también en las otras. Esto puede reflejar tanto la calidad educativa como factores socioeconómicos más amplios.

Estos hallazgos servirán como punto de partida para análisis más complejos, especialmente en etapas posteriores donde se crucen variables entre conjuntos para detectar correlaciones o diferencias significativas entre zonas o poblaciones.

6. Reporte de calidad de datos

Como se mencionó anteriormente, antes de avanzar con el análisis exploratorio y la fase de transformación, se realizó una evaluación de la calidad de los datos disponibles. Esta revisión tuvo como propósito identificar problemas de integridad, consistencia, completitud y tipificación incorrecta que pudieran interferir con los procesos analíticos posteriores.

La validación de calidad incluyó tanto técnicas cuantitativas (conteo de nulos, duplicados, detección de valores no válidos) como observaciones cualitativas basadas en los tipos de datos de cada conjunto. En terminos generales, el reporte se organizó en dos fases principales: análisis de valores faltantes y propuesta de tratamiento para las columnas afectadas.

6.1. Análisis de valores faltantes

Para cada uno de los conjuntos de datos seleccionados se realizó un diagnóstico inicial de calidad basado en el conteo de valores nulos, vacíos y duplicados. Esta revisión permitió identificar columnas potencialmente problemáticas, con el fin de definir estrategias de tratamiento en las etapas posteriores.

NYPD Arrest Data.
Para este conjunto ólo se detectaron valores nulos en dos columnas clave: LAW_CAT_CD (categoría legal del delito), con 687 registros nulos, lo que representa aproximadamente el 0,48 %, y KY_CD (código estandarizado del delito), con 10 registros nulos, equivalentes al 0,007 El resto de las columnas no contiene valores faltantes ni registros duplicados, por lo que se considera un conjunto de alta calidad para el análisis.

NYCgov Poverty Measure Data.
Este conjunto presenta un mayor grado de incompletitud. Se identificaron valores nulos en 61 columnas, principalmente relacionadas con variables laborales y de transporte. Por ejemplo, ENG (nivel de inglés en el hogar) presenta un 55,01 % de nulos, JWTR (medio de transporte al trabajo) un 51,20 %, y WKW (semanas trabajadas durante el año) un 45,98 %. Sin embargo, las variables ajustadas por ingreso, como SEMP_adj (ingreso por trabajo por cuenta propia), SSIP_adj (ingreso por seguridad suplementaria) y WAGP_adj (ingreso por salario), no contienen valores faltantes. Tampoco se detectaron registros duplicados. El tratamiento posterior deberá enfocarse en filtrar o imputar columnas con alta proporción de nulos, priorizando aquellas con mayor relevancia para el modelado.

Motor Vehicle Collisions – Vehicles.
Este conjunto, el cual contiene más de 4,4 millones de registros, presenta un nivel considerable de datos faltantes. Entre las columnas más afectadas se encuentran PUBLIC_PROPERTY_DAMAGE_TYPI (tipo de daño a propiedad pública), con un 99,3 % de valores faltantes, VEHICLE_MODEL (modelo del vehículo) con un 98,8 %, y VEHICLE_DAMAGE_3 (zona secundaria de daño) con un 76,7 %. Otras columnas como DRIVER_LICENSE_STATUS (estado de la licencia del conductor) y DRIVER_SEX (sexo del conductor) presentan vacíos por encima del 50 %. A pesar de ello, columnas fundamentales como COLLISION_ID (identificador del accidente), CRASH_DATE (fecha del accidente) y CRASH_TIME (hora del accidente) no contienen nulos ni duplicados. Debido al tamaño del conjunto, será necesario aplicar filtros y reducir dimensionalidad para facilitar su análisis

SAT NYC.
Si bien para este conjunto no se encontraron valores nulos explícitos, se detectó que las columnas Num_of_SAT_Test_Takers (número de estudiantes que presentaron el examen), SAT_Critical_Reading_Avg_Score (puntaje promedio en lectura crítica), SAT_Math_Avg_Score (puntaje promedio en matemáticas) y SAT_Writing_Avg_Score (puntaje promedio en escritura) fueron interpretadas como texto. Esto indica la presencia de valores no numéricos , lo que será abordado más adelante como parte del análisis semántico.

En resumen, únicamente el conjunto NYPD Arrest Data presenta una estructura completa y lista para el análisis inmediato. Los conjuntos NYCgov Poverty Measure Data y Motor Vehicle Collisions – Vehicles requieren un tratamiento más profundo de los valores faltantes, dada la alta proporción de vacíos en múltiples columnas. Por su parte, el conjunto SAT NYC requiere un proceso de depuración específico para normalizar las columnas que fueron interpretadas como texto y reemplazar los valores no numéricos identificados.

6.2. Detección de valores no numéricos en columnas numéricas

Además de los valores nulos explícitos, se realizó una inspección de aquellas columnas que deberían contener únicamente datos numéricos. Este paso fue necesario debido a que algunos conjuntos presentaban columnas con tipo de dato texto (`string`) a pesar de representar variables cuantitativas. El objetivo fue identificar valores no numéricos o codificaciones erróneas que impidieran una conversión correcta.

En los conjuntos NYPD Arrest Data, NYCgov Poverty Measure Data y Motor Vehicle Collisions – Vehicles se evaluaron todas las columnas que representan variables numéricas, con el fin de identificar posibles valores no válidos o inconsistencias de tipo. En los tres casos se confirmó que las variables numéricas están correctamente tipadas y no presentan valores alfabéticos, símbolos extraños ni codificaciones inválidas. Si bien algunas variables incluyen valores negativos o ceros, estos son coherentes con el significado de cada atributo y no requieren tratamiento adicional en esta etapa.

SAT NYC. En este conjunto se identificaron valores no válidos del tipo texto en todas las columnas que deberían ser numéricas: Num_of_SAT_Test_Takers (número de estudiantes que presentaron el SAT), SAT_Critical_Reading_Avg_Score (puntaje promedio en lectura crítica),

SAT_Math_Avg_Score (puntaje promedio en matemáticas) y SAT_Writing_Avg_Score (puntaje promedio en escritura). En cada una de estas columnas se detectó la presencia del carácter 's' como valor no numérico. Esto impide la conversión automática a tipo entero y representa un caso de falsos nulos, por lo que requerirá un proceso de limpieza que reemplace dichos valores por `null` antes de realizar cualquier análisis cuantitativo.

Con esto, se confirma que solo uno de los cuatro conjuntos de datos (SAT NYC), requiere una corrección directa de tipo para lograr su análisis numérico. Los demás conjuntos presentan una estructura numérica válida, sin presencia de datos atípicos o no numéricos en sus variables clave.

6.3. Propuesta de tratamiento

Con base en lo encontrado durante el análisis de calidad de datos, se plantean las siguientes estrategias de tratamiento para preparar los conjuntos antes de su exploración y modelado.

NYPD Arrest Data. Dado que este conjunto presenta únicamente una proporción menor de valores nulos en las columnas LAW_CAT_CD (categoría legal del delito) y KY_CD (código estandarizado del delito), se propone realizar una eliminación directa de los registros incompletos en estas variables, ya que representan menos del 1 % del total. No se requieren imputaciones ni transformaciones adicionales.

NYCgov Poverty Measure Data. Este conjunto presenta múltiples columnas con valores faltantes, especialmente en variables relacionadas con condiciones laborales y transporte. Se propone:

- Eliminar columnas con más del 50 % de nulos.
- Aplicar imputación por moda o mediana en variables numéricas o por la categoría `UNKNOWN` en variables categóricas.
- Mantener las variables económicas que no presentan nulos.

Motor Vehicle Collisions – Vehicles. Debido al alto volumen de datos y al gran número de columnas con vacíos, se plantea:

- Filtrar columnas con más del 50 % de valores nulos.
- Reemplazar valores nulos en campos categóricos con la etiqueta `UNKNOWN`..
- Mantener las columnas completas.

SAT NYC Este conjunto requiere una transformación específica en sus variables cuantitativas, que actualmente están representadas como cadenas de texto. Se propone el siguiente tratamiento:

- Reemplazar los valores s por null en las columnas de puntajes y número de estudiantes, ya que impiden el análisis numérico. - Convertir dichas columnas a tipo numérico, con el propósito de poder realizar operaciones estadísticas. - Una vez realizado el casteo, se repetirá el proceso de conteo de valores nulos y duplicados con el nuevo esquema de datos.

.

7. Planteamiento de preguntas sobre los datos

En esta sección se formulan una serie de preguntas clave orientadas al análisis territorial y social de la ciudad de Nueva York, a partir de los conjuntos de datos previamente explorados. Estas preguntas no solo reflejan patrones identificados en arrestos, accidentes, condiciones socioeconómicas y desempeño académico, sino que buscan generar hallazgos que puedan traducirse en acciones concretas por parte del equipo de gobierno. Todas las preguntas fueron diseñadas con un enfoque en el impacto social, la desigualdad territorial y la formulación de estrategias de prevención y mejora de calidad de vida.

7.1. Preguntas principales

1. ¿Qué tipos de delitos son más frecuentes en zonas de bajos ingresos, y qué diferencias hay respecto a zonas de mayor ingreso?
2. ¿Hay relación entre los puntajes promedio del SAT por distrito y la tasa de arrestos?
3. ¿Cuáles son los barrios o distritos con mayor concentración de delitos violentos?
4. ¿En qué horarios o temporadas se concentran más accidentes?
5. ¿Qué relación existe entre el nivel educativo promedio y la condición de pobreza en los distintos distritos?
6. ¿Las zonas con mayores niveles de pobreza presentan también mayores tasas de arrestos? ¿Cómo varía esta relación entre distritos?
7. ¿Qué tipos de vehículos están más involucrados en colisiones con daño a propiedad pública?
8. ¿En qué días de la semana se cometen más delitos violentos?

7.2. Justificación de su relevancia

Cada una de las preguntas anteriores tiene el potencial de generar valor estratégico para el equipo de gobierno, al permitir una mejor comprensión del territorio y la orientación de políticas públicas más efectivas. A continuación, se presenta la justificación de cada una:

- **Pregunta 1:** Permite identificar patrones delictivos asociados a condiciones económicas, lo cual es fundamental para diseñar estrategias de seguridad diferenciadas según el contexto socioeconómico.
- **Pregunta 2:** Analizar la relación entre desempeño académico y arrestos podría revelar dinámicas de exclusión educativa y social, y abrir el camino a intervenciones escolares con enfoque preventivo.
- **Pregunta 3:** Conocer los distritos con mayor cantidad de delitos violentos ayuda a identificar territorios prioritarios para la intervención estatal, permitiendo enfocar estrategias de prevención y asignación de recursos en las zonas con mayor incidencia.

- **Pregunta 4:** Identificar horarios o temporadas con mayor concentración de incidentes ayuda a planificar mejor la asignación de recursos policiales y de tránsito.
- **Pregunta 5:** Explorar la relación entre educación y pobreza ofrece insumos para políticas públicas de inclusión educativa y reducción de desigualdad.
- **Pregunta 6:** Busca comprender cómo el nivel educativo se relaciona con las condiciones de pobreza en los distintos distritos, lo que puede ayudar a identificar brechas de oportunidad y orientar programas educativos o sociales hacia las comunidades más vulnerables.
- **Pregunta 7:** Facilita la identificación de los tipos de vehículos más problemáticos desde una perspectiva de seguridad urbana y daño a infraestructura pública, lo que puede derivar en regulaciones más estrictas o campañas de prevención.
- **Pregunta 8:** Analizar los días de la semana con mayor cantidad de delitos violentos permite optimizar la planificación operativa de las autoridades, ajustando los esfuerzos de patrullaje y respuesta en función de los picos de actividad delictiva.

8. Transformaciones, filtrado y limpieza inicial

Una vez finalizada la etapa de evaluación de calidad de los datos, se procedió con el desarrollo de un conjunto de transformaciones, filtros y limpiezas iniciales orientadas a garantizar la consistencia, completitud y utilidad de las variables involucradas. Esta fase tuvo como propósito preparar cada conjunto de datos para su análisis posterior, eliminando inconsistencias, depurando registros incompletos y normalizando formatos. Las acciones realizadas se guiaron por la propuesta de tratamiento definida previamente y se según las características propias de cada conjunto. Se buscó intervenir lo mínimo necesario para conservar la mayor cantidad posible de información relevante.

8.1. Transformaciones preliminares

Durante esta etapa se realizaron transformaciones básicas sobre los conjuntos de datos, con el objetivo de dejar las variables clave listas para su análisis posterior. Estas transformaciones incluyeron cambios de tipo, creación de nuevas columnas y ajustes en el formato de ciertos valores que venían mal representados.

En el caso del conjunto SAT NYC, algunas columnas que deberían contener números estaban registradas como texto debido a la presencia del carácter `s`, que indicaba falta de reporte. Se reemplazaron estos valores por nulos, se convirtieron las columnas a tipo numérico, y se creó una nueva columna que suma los puntajes por área. Luego de esta transformación, se eliminaron 57 registros que no contenían ningún dato útil, quedando un total de 421 observaciones válidas.

En el conjunto de pobreza se eliminaron columnas con más del 50 % de nulos y se imputaron los valores faltantes en las variables restantes. Para las columnas numéricas

se usó la mediana, y para las categóricas una etiqueta genérica unknown. El resultado fue una tabla sin datos faltantes y con 59 columnas útiles.

El conjunto de colisiones vehiculares fue tratado de forma similar. Se eliminaron las columnas más incompletas y se imputaron los valores faltantes en las variables restantes. Además, se unificaron los distintos valores que representaban ausencia de información (como "Unspecified", "N/A", "Unknown", entre otros) bajo una sola categoría estándar: unknown. En total se conservaron 17 columnas, y no fue necesario eliminar filas, ya que la limpieza se centró en estandarizar y completar los valores existentes.

Por último, en el conjunto de arrestos se encontraron nulos en dos columnas clave que definen el tipo de delito. Como estas variables son esenciales y los valores faltantes eran pocos (menos del 0.5 %), se decidió eliminar directamente esas filas. El conjunto final quedó con 142.797 registros completos.

8.2. Filtrados aplicados

En esta etapa se aplicaron filtros sobre los conjuntos, con el objetivo de enfocar el análisis en registros relevantes y reducir información poco útil o poco representativa, sin comprometer la cobertura general de los datos. Se mantuvo el principio de intervenir lo menos posible, filtrando únicamente cuando era necesario y justificado.

En el caso del conjunto de arrestos, todos los registros correspondían al año 2025, por lo que no fue necesario filtrar por fechas. Sin embargo, se consideró importante restringir el análisis a delitos de mayor gravedad. Por esta razón, se conservaron únicamente los registros cuya categoría legal (LAW_CAT_CD) correspondía a felonías (F) o delitos menores (M), excluyendo las violaciones menores (V). Este filtro redujo el tamaño de la base de 142.797 a 140.086 registros, lo cual representa una pérdida menor al 2 % del total, pero permite centrar el análisis en eventos de mayor impacto social y territorial.

Para los demás conjuntos de datos —pobreza, colisiones vehiculares y SAT NYC— no se aplicaron filtros adicionales en esta etapa, ya que presentaban una cobertura amplia, datos recientes o bien definidos, y variedad suficiente. Se consideró que aplicar filtros adicionales podría reducir innecesariamente el volumen de información disponible sin aportar un valor claro al análisis.

8.3. Limpiezas realizadas

En términos generales, los conjuntos de datos fueron sometidos a un proceso de limpieza enfocada en mejorar su calidad y garantizar su utilidad analítica. Este proceso incluyó la eliminación de columnas con altos porcentajes de valores nulos, algunos filtros específicos orientados a mejorar el enfoque del análisis, y transformaciones necesarias para corregir formatos, normalizar tipos de datos y crear variables clave. Además, en ciertos conjuntos se realizó una estandarización de valores categóricos, unificando distintas expresiones de datos no informados bajo una misma categoría. Se procuró intervenir lo menos posible, aplicando solo las acciones indispensables para asegurar la consistencia, completitud y coherencia de la información disponible.

9. Web scraping de datos poblacionales

Para cumplir con el requerimiento de realizar un proceso de web scraping sobre la población de Nueva York, se intentó inicialmente acceder al enlace oficial proporcionado por el enunciado del proyecto:

`https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoodpop.htm`

Sin embargo, este enlace actualmente se encuentra fuera de servicio (error 404), por lo que se optó por otra alternativa: el uso de la API oficial del U.S. Census Bureau, específicamente la del *American Community Survey (ACS) 5-Year 2020*.

Mediante esta API, se consultó la variable `B01003_001E`, correspondiente al total de población por condado. En el caso de la ciudad de Nueva York, los condados consultados corresponden a sus cinco distritos: Bronx, Brooklyn (Kings), Manhattan (New York), Queens y Staten Island (Richmond). Se usaron sus respectivos códigos FIPS para realizar la consulta, y posteriormente se construyó un *DataFrame* con la información.

El resultado de la consulta fue el siguiente:

- **Brooklyn (Kings)**: 2,576,771 habitantes
- **Queens**: 2,270,976 habitantes
- **Manhattan (New York)**: 1,629,153 habitantes
- **Bronx**: 1,427,056 habitantes
- **Staten Island (Richmond)**: 475,596 habitantes

A partir de estos datos se generó la siguiente visualización:

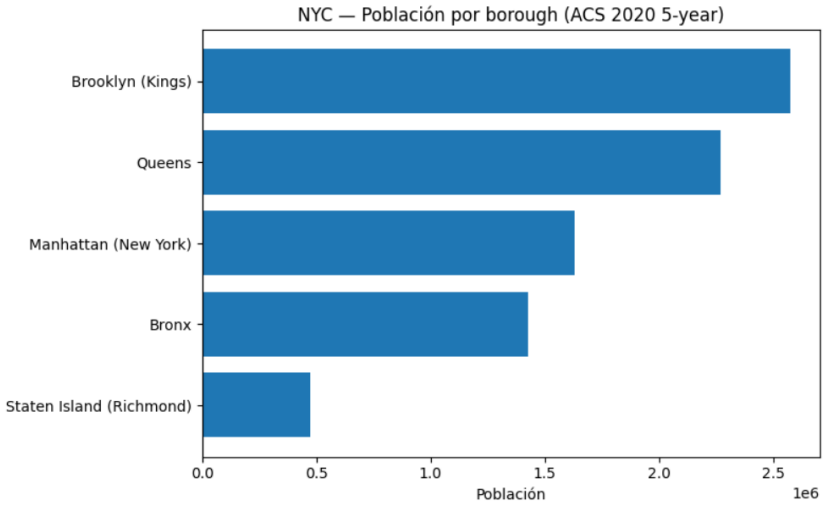

Figura 21: Distribución poblacional por distrito en la ciudad de Nueva York (ACS 2020 5-Year)

Este gráfico permite visualizar cómo se distribuye la población en los diferentes distritos de la ciudad. Brooklyn y Queens destacan como las zonas más densamente pobladas, mientras que Staten Island presenta una concentración significativamente menor.

Esta información será fundamental en etapas posteriores del análisis, ya que permitirá normalizar indicadores como arrestos o accidentes por cada 100,000 habitantes y hacer comparaciones más equitativas entre zonas con diferentes tamaños poblacionales.

10. Consulta climática con OpenWeatherMap

Como parte del segundo bono opcional de la entrega, se realizó una consulta a la API de OpenWeatherMap utilizando el endpoint de pronóstico a 5 días con cortes cada 3 horas. La ciudad consultada fue *New York, US*, y se extrajeron variables como temperatura, humedad, precipitación (lluvia y nieve), velocidad del viento y el estado general del clima.

La respuesta fue exitosa (status 200), y los datos fueron procesados para construir un *DataFrame* ordenado cronológicamente. En las primeras observaciones ya se evidencian patrones relevantes: cielos despejados durante el 21 de octubre, seguidos de episodios de lluvia ligera durante la madrugada del día 22, con temperaturas que oscilan entre los 19°C y los 11°C.

Se generaron dos visualizaciones a partir de esta información. La primera muestra la evolución de la temperatura durante los próximos días junto con las precipitaciones registradas en cada intervalo de 3 horas. Esta visualización permite identificar los ciclos térmicos diarios, así como los momentos de mayor precipitación. Por ejemplo, el 22 de octubre se concentran los valores más altos de lluvia, con varios picos cercanos a los 2 mm por bloque de 3 horas.

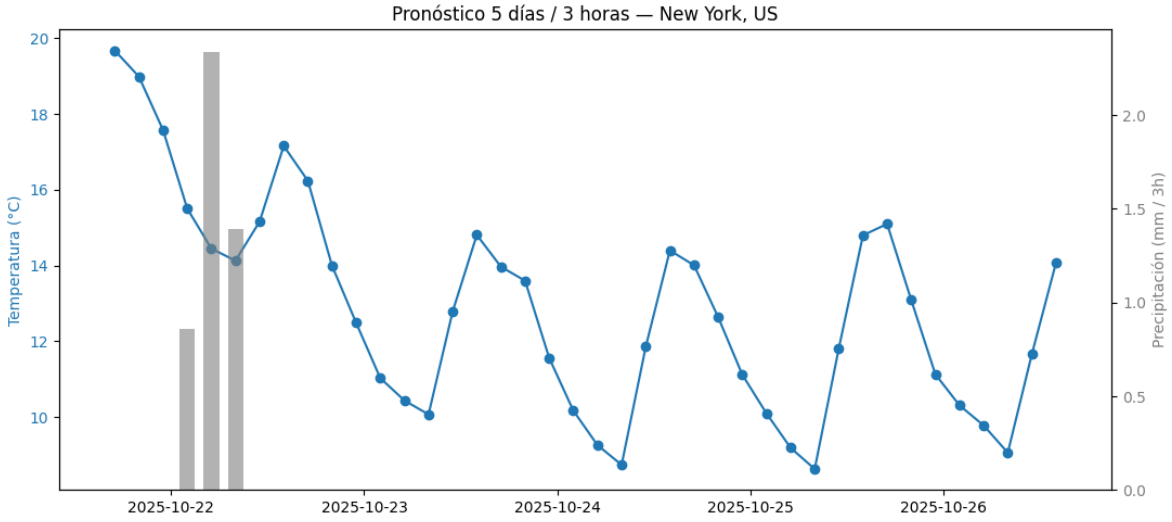

Figura 22: Pronóstico de temperatura y precipitación por intervalos de 3 horas — New York, US

La segunda visualización agrega la información a nivel diario, mostrando la temperatura máxima y mínima por jornada, así como la lluvia total diaria acumulada. En esta figura se observa un descenso progresivo de la temperatura a lo largo del periodo, comenzando cerca a los 20°C y descendiendo hasta valores cercanos a 14°C. La única jornada con lluvia significativa fue el 22 de octubre, con una acumulación superior a los 4.5 mm, mientras que el resto de los días se mantuvieron secos según el pronóstico.

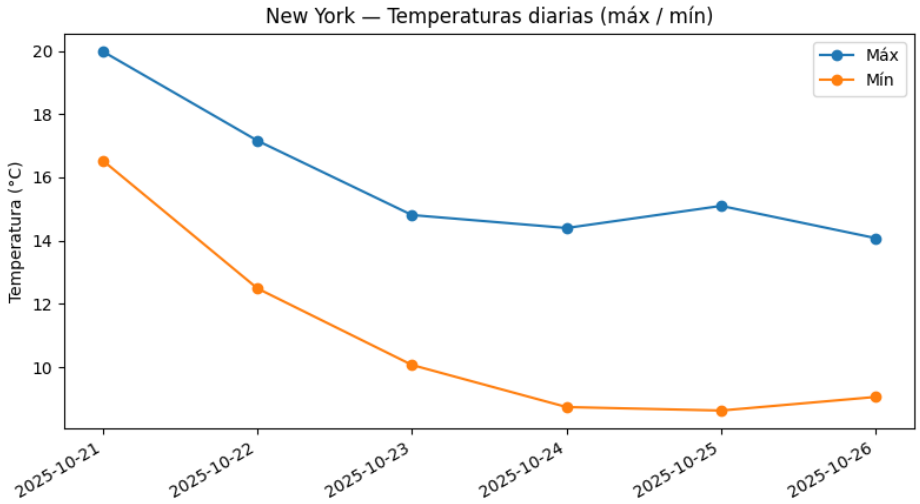

Figura 23: Temperaturas máximas y mínimas diarias — New York, US

Estas visualizaciones pueden servir más adelante para analizar posibles relaciones entre condiciones climáticas y fenómenos como la movilidad o la accidentalidad vial en la ciudad.

11. Conclusiones y recomendaciones de la primera fase

Durante esta primera entrega se logró consolidar el entendimiento del negocio, la recopilación y descripción técnica de los conjuntos de datos, así como su exploración inicial mediante análisis estadísticos y visualizaciones. Estos avances permitieron obtener una comprensión integral del contexto urbano y social de la ciudad de Nueva York, identificando los principales retos asociados a la seguridad, la movilidad y las condiciones socioeconómicas de su población.

El procesamiento distribuido en el clúster de Apache Spark permitió validar la infraestructura propuesta y garantizar la capacidad para manejar grandes volúmenes de información de manera eficiente. A su vez, la revisión de calidad de datos permitió detectar problemas comunes como valores nulos, registros no numéricos y categorías inconsistentes, los cuales fueron tratados mediante procesos de limpieza y transformación que dejaron las bases listas para su análisis posterior.

En términos de resultados, se evidenciaron patrones relevantes: la concentración de arrestos en hombres adultos jóvenes, la mayor siniestralidad vial asociada a la distracción del conductor, la persistente desigualdad económica entre distritos y la correlación positiva entre los puntajes académicos del SAT. Estos hallazgos preliminares sientan las bases para los análisis cruzados y modelamientos que se desarrollarán en la segunda entrega.

12. Segunda Entrega – Modelado y Análisis Avanzado

Esta segunda entrega corresponde a la etapa final del proyecto, en la cual se completan los procesos de preparación de datos y se desarrollan los análisis que permitirán responder las preguntas de negocio planteadas. A partir del trabajo exploratorio realizado en la fase anterior, esta parte del informe se enfoca en generar hallazgos integrados, aplicar técnicas de modelado y proponer conclusiones y recomendaciones con base en la evidencia encontrada.

12.1. Recapitulación de la fase 1

Durante la primera entrega del proyecto se consolidaron las etapas iniciales del ciclo CRISP-DM, centradas en el entendimiento del negocio y los datos. El objetivo principal fue construir una base de conocimiento sobre la ciudad de Nueva York y sus problemáticas sociales prioritarias, en particular la alta tasa de arrestos y la cantidad de accidentes viales.

Para ello, se seleccionaron y procesaron cuatro conjuntos de datos oficiales de Open Data NYC:

- **NYPD Arrest Data (2025):** Incluye más de 140 mil registros detallando edad, sexo, tipo de delito, raza y ubicación del arresto.
- **Motor Vehicle Collisions – Vehicles:** Con más de 4 millones de registros sobre accidentes viales, tipo de vehículo, daños, condiciones del conductor y factores contribuyentes.
- **NYCgov Poverty Measure Data:** Contiene indicadores socioeconómicos como tasa de pobreza, ingreso mediano y distribución por distrito.
- **SAT NYC:** Resultados agregados por escuela en lectura crítica, matemáticas y escritura, usados como proxy educativo.

A nivel técnico, se implementó un **clúster Apache Spark** configurado manualmente sobre Rocky Linux, con tres nodos (1 master + 2 workers), lo que permitió realizar el procesamiento distribuido desde un entorno Jupyter con PySpark..

Posteriormente, se desarrolló una exploración estadística y visual de cada conjunto. Entre los hallazgos clave destacan:

- Alta concentración de arrestos en hombres entre 25 y 44 años, con predominancia de delitos menores en distritos como Brooklyn y el Bronx.
- Más del 60 % de los registros de accidentes presentan causas "no informadas"; los factores más frecuentes fueron distracción y no ceder el paso.
- El distrito del Bronx presentó la tasa de pobreza más alta (˜26 %), mientras que Staten Island y Manhattan registraron las más bajas.

También se abordó un análisis de calidad de datos, identificando columnas con altos niveles de nulos y valores no válidos (especialmente en SAT y colisiones). Se aplicaron técnicas de limpieza, imputación y estandarización, dejando las bases listas para modelado.

Finalmente, se formularon ocho preguntas de negocio, orientadas a entender relaciones entre pobreza, educación, criminalidad y movilidad. Estas preguntas guían la fase actual del proyecto.

12.2. Enfoque actual de la fase 2

Esta segunda fase del proyecto tiene como propósito completar el proceso analítico iniciado en la entrega anterior. Mientras la primera etapa se centró en entender el contexto, explorar los datos y formular preguntas clave, en esta fase se espera finalizar las tareas de limpieza, aplicar transformaciones más avanzadas y dar respuesta a las preguntas de negocio previamente planteadas.

El eje principal del trabajo estará en interpretar los datos de forma integrada, cruzando variables de distintos conjuntos y generando visualizaciones e indicadores que permitan explicar mejor las dinámicas observadas en la ciudad. A partir de este análisis, se buscará identificar patrones relevantes que sustenten hallazgos concretos y, con base en ellos, construir recomendaciones para el gobierno de Nueva York.

Además, se incorporarán técnicas de aprendizaje de máquina como complemento del análisis, aplicando modelos supervisados y no supervisados para enriquecer la comprensión de los datos.

En conjunto, esta fase representa el cierre del proyecto: pasar de los datos y su análisis a conclusiones y propuestas aplicables.

12.3. Objetivos específicos de esta etapa

Los objetivos de esta segunda fase están orientados a cerrar el ciclo analítico y generar conclusiones que respondan al propósito del proyecto. De forma específica, se busca:

- Finalizar los procesos de limpieza, transformación y estandarización de los conjuntos de datos utilizados.
- Responder, mediante análisis integrados y visualizaciones, las preguntas formuladas en la primera fase.
- Identificar patrones y correlaciones relevantes que permitan comprender mejor las dinámicas sociales, económicas y territoriales de Nueva York.
- Formular conclusiones a partir de los hallazgos observados.
- Proponer recomendaciones para el gobierno de la ciudad, enfocadas en mejorar indicadores como la seguridad ciudadana, la movilidad y la equidad social.
- Aplicar al menos una técnica de aprendizaje supervisado y una no supervisada, con el fin de complementar los análisis.
- Evaluar el desempeño de los modelos implementados y analizar su utilidad frente a los objetivos del proyecto.

13. Transformaciones y Filtros Finales

En esta etapa se consolidaron los filtros y transformaciones definitivas aplicadas a los distintos conjuntos de datos. El propósito fue dejar la información en un estado adecuado para responder las preguntas de negocio y preparar las variables que se utilizarán en los modelos de aprendizaje de máquina. A continuación se describen las depuraciones realizadas y las nuevas columnas generadas.

13.1. Recapitulación de transformaciones y filtros aplicados en la fase 1

Durante la primera fase, se realizaron transformaciones y limpiezas iniciales sobre los cuatro conjuntos de datos principales, con el objetivo de garantizar su consistencia y preparar su análisis posterior.

En el conjunto **SAT NYC**, se reemplazaron los valores *'s'* por nulos en las columnas de puntajes, se convirtieron a tipo numérico y se creó una nueva variable con el total del puntaje SAT. Posteriormente se eliminaron 57 registros sin datos válidos.

Para el conjunto de **pobreza**, se eliminaron las columnas con más del 50 % de valores nulos, se imputaron las variables numéricas con la mediana y las categóricas con la etiqueta *unknown*. El resultado fue una tabla depurada con 59 columnas útiles.

En el conjunto de **colisiones vehiculares**, se eliminaron columnas con alta proporción de vacíos y se unificaron los valores no informados bajo una categoría estándar. Se conservaron 17 columnas clave y no fue necesario eliminar registros.

En el conjunto de **arrestos**, se eliminaron directamente las filas con valores nulos en las columnas `LAW_CAT_CD` y `KY_CD`, dado que representaban menos del 0.5 % del total.

Adicionalmente, se aplicaron filtros para mejorar el enfoque analítico. En el conjunto de arrestos, se excluyeron las violaciones menores (código legal `V`), conservando únicamente delitos clasificados como felonías (`F`) y delitos menores (`M`), lo cual redujo el conjunto a 140.086 registros.

A continuación se presenta una tabla que resume los filtros y transformaciones aplicados en la fase 1:

Conjunto	**Cambios realizados**
SAT NYC	Reemplazo de valores no numéricos, casteo a tipo numérico, creación de puntaje total y eliminación de registros vacíos
Pobreza	Eliminación de columnas con muchos nulos, imputación por mediana y *unknown*
Colisiones vehiculares	Reducción de columnas, estandarización de valores no informados, sin eliminación de registros
Arrestos	Eliminación de registros con nulos clave; filtrado de violaciones menores (sólo se dejaron `F` y `M`)

Cuadro 1: Transformaciones y filtros aplicados en la fase 1

13.2. Filtros adicionales

Antes de aplicar nuevas transformaciones, se optó por realizar primero los filtros. Esto se hizo para evitar transformar y analizar datos que luego podrían ser eliminados, lo cual generaría inconsistencias o análisis sobre subconjuntos que no serían los definitivos. A continuación se explican los principales filtros aplicados:

- **Límite superior de ingreso:** En una revisión inicial, se detectaron registros con ingresos extremadamente altos, incluso cercanos al millón de dólares. Estos valores sesgaban la distribución y podían generar inconsistencias en los análisis posteriores. Por eso, se decidió eliminar los casos con ingresos ajustados superiores a $250.000, un umbral comúnmente usado en estudios sociales en Estados Unidos para definir hogares de ingresos muy altos. Tras aplicar este filtro, el número de registros pasó de 68.273 a 66.066, lo que permitió trabajar con una muestra más representativa del contexto urbano general.

- **Instituciones con pocos estudiantes SAT:** Se eliminaron los registros de colegios donde menos de 30 estudiantes presentaron el examen SAT. Esto se hizo para evitar que instituciones demasiado pequeñas influyeran de forma desproporcionada en los análisis, ya que un solo resultado atípico puede alterar el promedio cuando el tamaño de muestra es muy bajo. El umbral de 30 se eligió porque, según algunas teorías de probabilidad, a partir de ese punto es posible tratar la

muestra como representativa de una población, permitiendo comparaciones más estables entre instituciones. Tras aplicar este filtro, el conjunto se redujo de 421 a 359 registros.

- **Evaluación del filtro por daño a propiedad pública:** Se consideró aplicar un filtro para dejar solo los accidentes donde hubo daño a propiedad pública. Sin embargo, al ver que solo una pequeña fracción de los datos (17.328 casos de más de 4 millones) cumplía con esa condición, se decidió no aplicarlo para evitar perder demasiada información.

13.3. Transformaciones adicionales

Luego de aplicar los filtros correspondientes, se realizaron varias transformaciones sobre los conjuntos de datos con el fin de facilitar los análisis, mejorar la interpretación de ciertas variables y preparar la información para su posterior uso en modelos de aprendizaje de máquina. A continuación se describen las principales transformaciones aplicadas:

- **Promedio total SAT por institución:** Se creó la variable `SAT_Total_Avg_Score`, calculando el promedio entre las tres secciones del examen SAT (lectura crítica, matemáticas y escritura). A diferencia de la suma total utilizada en la primera entrega, esta nueva métrica representa mejor el rendimiento medio por estudiante, evitando que puntajes desbalanceados entre secciones generen interpretaciones equivocadas.

- **Clasificación de rendimiento SAT:** Con base en la variable anterior, se categorizaron los colegios según su desempeño promedio usando tres niveles: BAJO (menor a 450), MEDIO (entre 450 y 600) y ALTO (mayor a 600). Esta transformación, almacenada en la columna `SAT_Rendimiento`, permite comparar el rendimiento académico de forma más sencilla e interpretar su relación con otras variables del proyecto.

- **Normalización del puntaje SAT:** Se creó la variable `SAT_Total_Norm` aplicando una normalización min–max, es decir, ajustando los puntajes promedio SAT para que quedaran dentro de un rango entre 0 y 1. Esta escala facilita las comparaciones entre escuelas y permite usar esta variable más adelante sin que los valores grandes o pequeños dominen sobre otros indicadores. En esta nueva versión, los números no representan un puntaje real, sino la posición de cada institución dentro del conjunto, donde 0 corresponde al puntaje más bajo observado y 1 al más alto.

- **Clasificación del ingreso ajustado:** Se creó una nueva columna llamada `Income_Avg`, que agrupa el ingreso ajustado de cada persona en tres niveles: BAJO (menos de $30.000), MEDIO ($30.000 a $70.000) y ALTO (más de $70.000). Esta segmentación ayuda a comparar el comportamiento de distintos indicadores en función del

nivel socioeconómico, como por ejemplo la frecuencia de arrestos o el rendimiento escolar.

- **Normalización del ingreso:** Se generó la variable `NYCgov_Income_Norm` aplicando normalización min-max sobre los ingresos ajustados, luego de eliminar previamente los valores extremos. Esto permitió obtener una distribución más balanceada para los análisis y para el entrenamiento de modelos de machine learning. Por ejemplo, ingresos como \$81.000 y \$117.000 se normalizaron a valores intermedios (0.42 y 0.54), mientras que los más altos quedaron cercanos a 1.0.
- **Día de la semana del accidente:** A partir del campo `CRASH_DATE` se extrajo el día de la semana en que ocurrió cada accidente, almacenado en la columna `CRASH_DAY`. Esta transformación facilita la exploración de patrones temporales en la accidentalidad, como la comparación entre fines de semana y días laborales.

13.4. Tabla resumen de cambios realizados

Con el objetivo de facilitar la visualización general del trabajo realizado en esta etapa, se presenta a continuación una tabla que resume los principales cambios aplicados sobre los conjuntos de datos. Estos incluyen tanto filtros para depurar información como transformaciones para mejorar la interpretación, el análisis comparativo y la preparación para modelos de aprendizaje automático.

Cambio realizado	**Descripción**
Filtro por ingreso	Se eliminaron registros con ingresos superiores a \$250.000 para evitar sesgos.
Filtro por tamaño muestral SAT	Se descartaron instituciones con menos de 30 estudiantes que presentaron el examen.
Evaluación de daño a propiedad pública	Se descartó aplicar el filtro por pérdida excesiva de información.
Promedio SAT por institución	Se creó una variable con el promedio entre las tres secciones del SAT.
Clasificación de rendimiento SAT	Se agruparon las instituciones en niveles: bajo, medio y alto.
Normalización SAT	Se escaló el puntaje promedio SAT a un rango de 0 a 1.
Clasificación del ingreso	Se clasificaron los ingresos en niveles: bajo, medio y alto.
Normalización del ingreso	Se aplicó normalización min–max sobre los ingresos ajustados.
Día del accidente	Se extrajo el día de la semana a partir de la fecha del accidente.

Cuadro 2: Resumen de filtros y transformaciones aplicados en la etapa 2

14. Respuesta a las Preguntas de Negocio

Una vez completadas las etapas de limpieza, transformación y análisis exploratorio de los datos, se procedió a responder las preguntas de negocio planteadas en la primera entrega del proyecto. Cada una de ellas busca dar sentido a los patrones observados en los conjuntos de datos de arrestos, accidentes, pobreza y educación en la ciudad de Nueva York, relacionando los resultados obtenidos con el contexto social y urbano de la ciudad.

En esta sección se presentan los principales hallazgos obtenidos a partir del procesamiento de los datos en el clúster de Spark, acompañados de visualizaciones y análisis interpretativos. Cada subsección aborda una de las preguntas formuladas previamente, explicando la metodología aplicada, los resultados obtenidos y las conclusiones más relevantes que pueden orientar la toma de decisiones por parte de las autoridades locales.

14.1. Relación entre tipo de delito e ingreso por zona

Para explorar si existe alguna relación entre el nivel de ingresos y los tipos de delitos más comunes en la ciudad, se comenzó analizando la variable `Income_Avg` del conjunto `df_pobreza`. Esta variable clasifica a la población en tres niveles de ingreso: bajo (menos de $30.000), medio (entre $30.000 y $70.000) y alto (más de $70.000). Con esta información se elaboró una tabla que muestra la proporción de personas en cada nivel por distrito, lo que permitió identificar qué zonas presentan mayores diferencias económicas.

El análisis de la tabla mostró que el Bronx concentra la mayor proporción de habitantes con ingresos bajos, mientras que Staten Island es el distrito con el porcentaje más alto de personas con ingresos altos. En el Bronx, cerca del 27 por ciento de la población pertenece al nivel bajo y solo un 27 por ciento al nivel alto, mientras que en Staten Island más de la mitad de los habitantes se ubica en el nivel alto y apenas un 15 por ciento en el nivel bajo. Con base en estos resultados, se eligieron estos dos distritos como casos representativos para comparar los tipos de delitos predominantes según el contexto económico: el Bronx como zona de ingresos más bajos y Staten Island como zona de ingresos más altos.

A partir de esta selección se agruparon los arrestos registrados en cada distrito por tipo de delito, usando la variable `OFNS_DESC`, y se contaron las apariciones para obtener los diez delitos más frecuentes. De esta manera fue posible identificar patrones distintos entre ambas zonas.

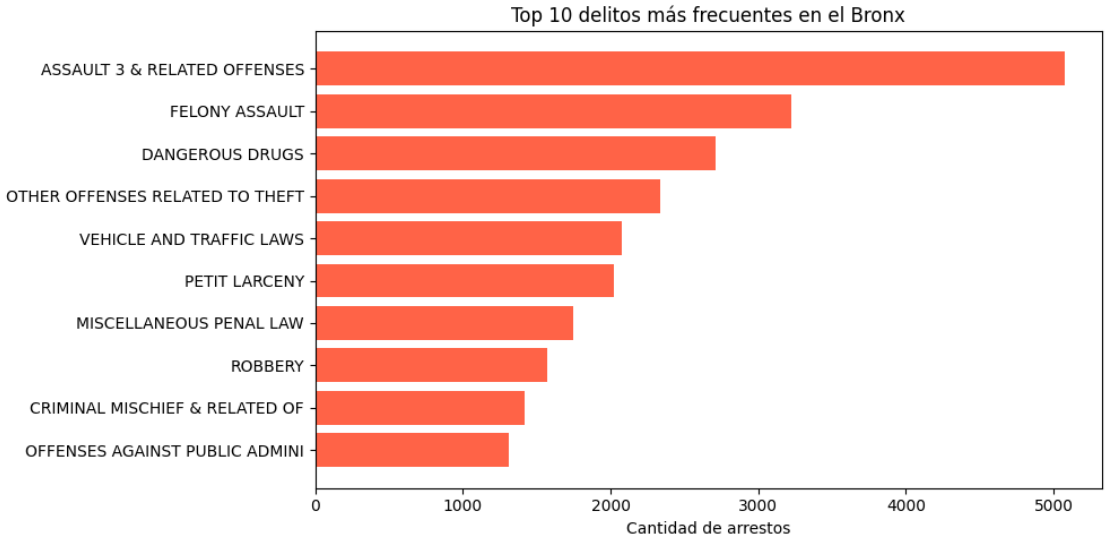

Figura 24: Top 10 delitos más frecuentes en el Bronx

En el Bronx se observó que los delitos más comunes están relacionados con la violencia física y el contacto directo entre personas. El de mayor número de arrestos fue `ASSAULT 3 & RELATED OFFENSES`, con más de cinco mil casos, seguido de `FELONY ASSAULT` y `DANGEROUS DRUGS`. También aparecen con frecuencia infracciones de tránsito, hurtos menores, robos con violencia y daños a la propiedad. Este conjunto de delitos refleja un entorno donde predominan los conflictos personales, las agresiones y las situaciones de riesgo asociadas al consumo o tráfico de drogas.

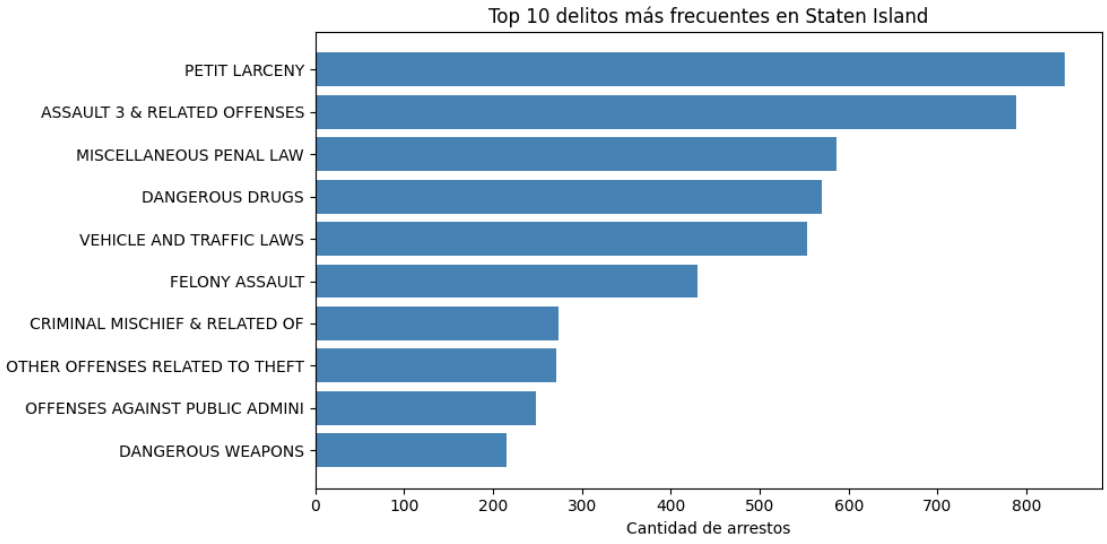

Figura 25: Top 10 delitos más frecuentes en Staten Island

En Staten Island, en cambio, el panorama es diferente. El delito más frecuente

fue `PETIT LARCENY`, relacionado con robos de bajo valor, seguido de `ASSAULT 3 & RELATED OFFENSES` y de `MISCELLANEOUS PENAL LAW`, que incluye infracciones menores como consumo de alcohol en vía pública o desorden en espacios abiertos. Otros delitos comunes fueron los relacionados con drogas, infracciones de tránsito y la posesión de armas, aunque en cantidades mucho menores que en el Bronx.

En general, los resultados muestran que en los distritos de menores ingresos tienden a predominar los delitos violentos y de contacto directo, mientras que en las zonas de mayores ingresos se registran más infracciones administrativas o de menor gravedad. Esto sugiere que las condiciones económicas y sociales pueden influir en el tipo de criminalidad que se presenta en cada zona, y que las estrategias de prevención deberían adaptarse a las características particulares de cada territorio.

14.2. Correlación entre SAT y tasa de arrestos por distrito

Para explorar la relación entre el nivel educativo y la criminalidad en la ciudad, se analizaron los datos del rendimiento promedio en el examen SAT y el número total de arrestos registrados en cada distrito de Nueva York. El propósito de este análisis fue identificar si los territorios con puntajes académicos más altos tienden a presentar una menor incidencia delictiva.

En primer lugar, se construyó una tabla con el puntaje promedio del SAT por distrito a partir del conjunto de datos `df_educacion`. Dado que el archivo no incluía directamente la información del distrito, fue necesario extraerla del código `DBN`, donde la letra inicial identifica el borough correspondiente (por ejemplo, "M" para Manhattan o "X" para el Bronx). A partir de esta codificación se agruparon las escuelas por distrito y se calcularon dos indicadores: el promedio general del SAT y la cantidad de instituciones reportadas en cada zona.

Los resultados muestran que Staten Island tiene el promedio más alto, con 460 puntos, mientras que el Bronx presenta el más bajo, con 383. Entre ambos extremos se ubican Queens (427), Manhattan (426) y Brooklyn (394). Estas diferencias reflejan una distribución educativa desigual entre distritos, tanto en número de instituciones como en desempeño promedio.

Posteriormente, se calcularon los arrestos totales por distrito utilizando la columna `ARREST_BORO` del conjunto `df_arrestos`. Los códigos de distrito fueron traducidos a sus nombres completos y se agruparon para obtener la cantidad de arrestos por territorio. De este modo, fue posible combinar ambos resultados —educación y criminalidad— en una sola tabla comparativa.

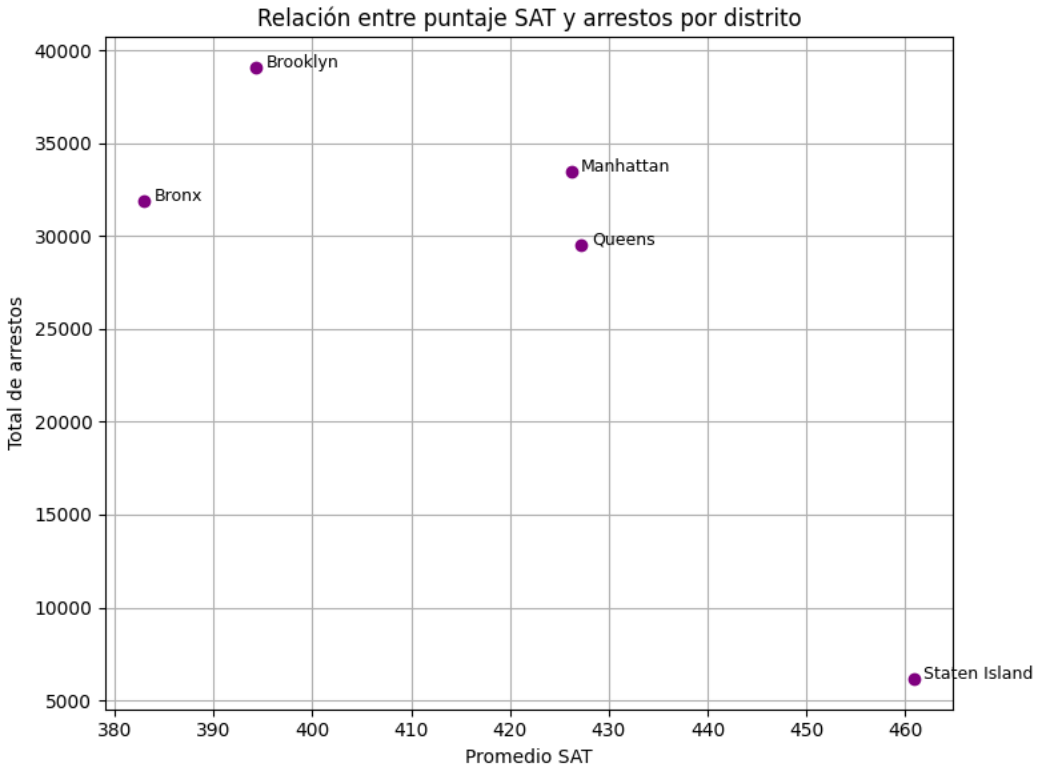

Figura 26: Relación entre puntaje SAT y arrestos por distrito

La gráfica anterior muestra la relación entre el promedio del SAT (eje horizontal) y el total de arrestos (eje vertical). Se observa una tendencia general donde los distritos con menores puntajes académicos presentan un mayor número de arrestos. Brooklyn y el Bronx, por ejemplo, tienen los valores más bajos en el SAT y, al mismo tiempo, las cifras más altas de arrestos. En cambio, Staten Island destaca con el puntaje más alto y la menor cantidad de arrestos registrados.

Aunque la relación no es perfectamente lineal, los resultados sugieren una correlación inversa entre nivel educativo y criminalidad: los distritos con mejor desempeño académico tienden a registrar menos incidentes delictivos. Esto podría interpretarse como una señal de que el acceso a una educación de mayor calidad, o a entornos escolares más estables, está asociado con una menor incidencia de comportamientos delictivos a nivel territorial. En consecuencia, las estrategias de prevención y mejora educativa podrían tener un efecto indirecto pero relevante sobre la reducción del crimen en la ciudad.

14.3. Distritos con mayor incidencia de delitos violentos

Con el fin de identificar cómo se distribuyen los delitos violentos en la ciudad de Nueva York, se agruparon los arrestos según el distrito donde ocurrieron. Para ello, se utilizó la columna `ARREST_BORO`, la cual fue transformada para mostrar los

nombres completos de los cinco boroughs: Brooklyn, Manhattan, Bronx, Queens y Staten Island. Posteriormente, se filtraron los registros que corresponden a delitos considerados violentos, como asaltos, robos, violaciones y homicidios.

Una vez agrupados los datos, se contabilizó la cantidad de arrestos en cada distrito. Los resultados muestran que Brooklyn registra el mayor número de arrestos por delitos violentos (7.559), seguido por Manhattan (7.184) y el Bronx (7.124), con cifras muy similares entre sí. Queens ocupa el cuarto lugar con 6.502 arrestos, mientras que Staten Island presenta un número considerablemente menor, con 1.016 casos.

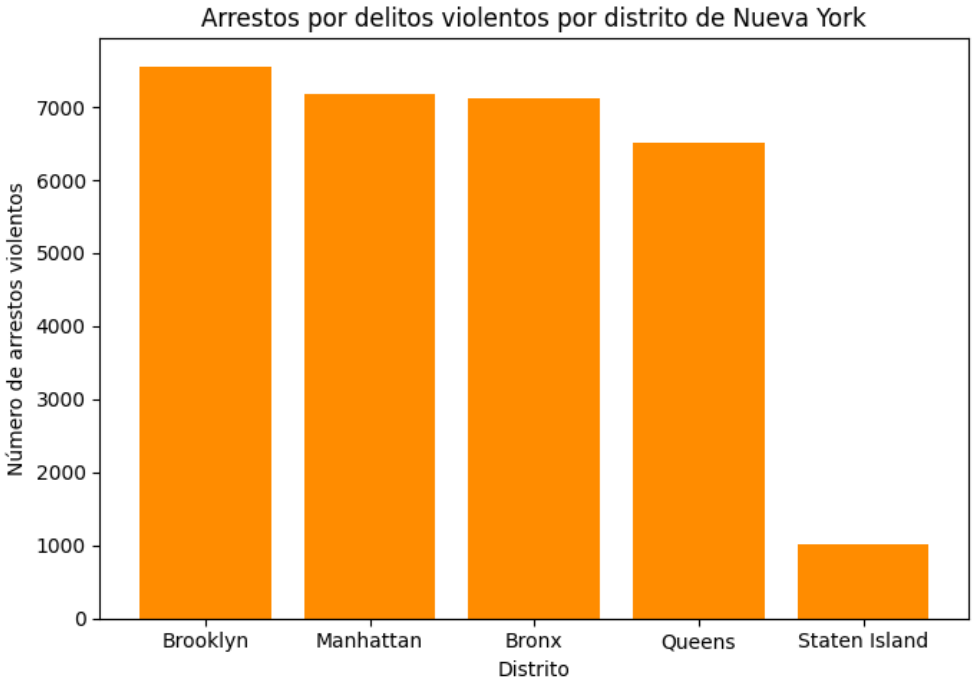

Figura 27: Arrestos por delitos violentos por distrito de Nueva York

La gráfica anterior permite visualizar con claridad la concentración de delitos violentos en las zonas más densamente pobladas y urbanizadas de la ciudad. Brooklyn, Manhattan y el Bronx destacan como los territorios donde la violencia es más frecuente, probablemente debido a factores asociados con la densidad poblacional, la actividad económica y las desigualdades sociales presentes en estas áreas.

Por el contrario, Staten Island se diferencia notablemente del resto, con una incidencia mucho menor de este tipo de delitos. Esto puede deberse a su menor tamaño poblacional, a un entorno urbano más residencial y a un nivel socioeconómico promedio más alto.

En conjunto, los resultados permiten concluir que la criminalidad violenta se concentra principalmente en los distritos centrales y de mayor población, lo que

refuerza la importancia de diseñar políticas de prevención y control que consideren las condiciones particulares de cada territorio.

14.4. Análisis temporal: ¿cuándo se concentran arrestos y accidentes?

Con el propósito de identificar en qué momentos de la semana ocurren con mayor frecuencia los accidentes de tránsito en Nueva York, se utilizó la variable `CRASH_DAY`, creada a partir de la fecha original del conjunto `df_accidentes`. Esta variable clasifica cada registro según el día de la semana en que se produjo el siniestro, desde lunes hasta domingo.

Los datos fueron agrupados por día y se calculó el total de accidentes registrados en cada uno. De esta forma, se construyó una gráfica que muestra la distribución semanal de los incidentes viales, permitiendo observar de manera visual los días de mayor y menor ocurrencia.

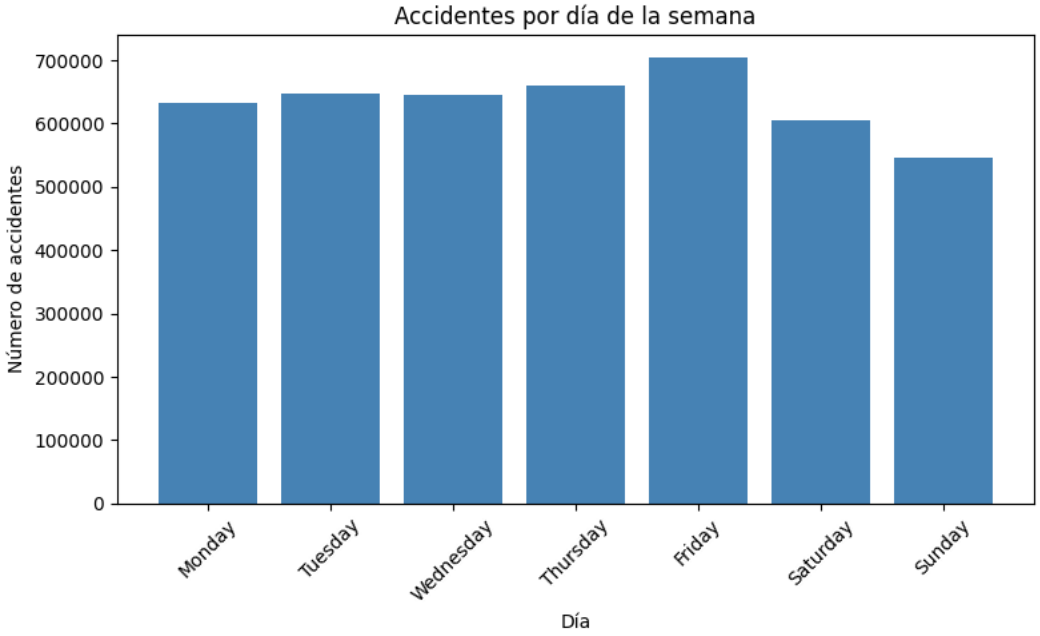

Figura 28: Accidentes por día de la semana en la ciudad de Nueva York

Los resultados muestran que los días laborales concentran la mayor parte de los accidentes, con un incremento progresivo hacia el final de la semana. El viernes es el día con más siniestros reportados, alcanzando un total de 705.286 casos, seguido del jueves (659.843) y del martes (648.294). Por el contrario, los fines de semana presentan los valores más bajos: el domingo con 545.471 y el sábado con 606.212 accidentes.

Este comportamiento puede explicarse por el aumento del tráfico en días hábiles, especialmente en horas de ingreso y salida laboral, así como por el cansancio acumulado y la mayor exposición al tránsito urbano. Los fines de semana, en cambio, la movilidad tiende a reducirse y las rutas suelen estar menos congestionadas, lo que se refleja en una menor cantidad de accidentes.

En conclusión, los datos evidencian que los accidentes de tránsito no se distribuyen de forma uniforme a lo largo de la semana, sino que se concentran principalmente en los días laborales, con un pico marcado los viernes. Este patrón puede resultar útil para orientar campañas de prevención vial y para planificar operativos de control en los momentos de mayor riesgo.

14.5. Nivel educativo promedio vs pobreza distrital

Para analizar la relación entre el nivel educativo promedio y la condición económica de los distintos distritos de Nueva York, se integraron dos fuentes de información: los resultados del examen SAT y los porcentajes de población clasificados por nivel de ingreso. Esta combinación permite observar si existen patrones entre el contexto socioeconómico y el desempeño académico en cada territorio.

En primer lugar, se tomó el puntaje promedio del SAT por distrito, calculado a partir de los registros escolares del conjunto `df_educacion`. Luego, se incorporaron los porcentajes de población con ingresos bajos y medios obtenidos del conjunto `df_pobreza`. De esta manera, se construyó una tabla comparativa con las siguientes variables: promedio SAT, número de escuelas, porcentaje de ingreso bajo y porcentaje de ingreso medio.

Los resultados muestran diferencias claras entre los distritos. Staten Island registra el promedio SAT más alto (460 puntos) y, al mismo tiempo, los menores porcentajes de población con ingresos bajos (14,9 %) y medios (30,9 %). En el extremo opuesto se encuentra el Bronx, con el puntaje promedio más bajo (383) y la mayor proporción de ingresos bajos (27,2 %) y medios (45,6 %). Brooklyn y Manhattan presentan valores intermedios, mientras que Queens se ubica ligeramente por encima del promedio general.

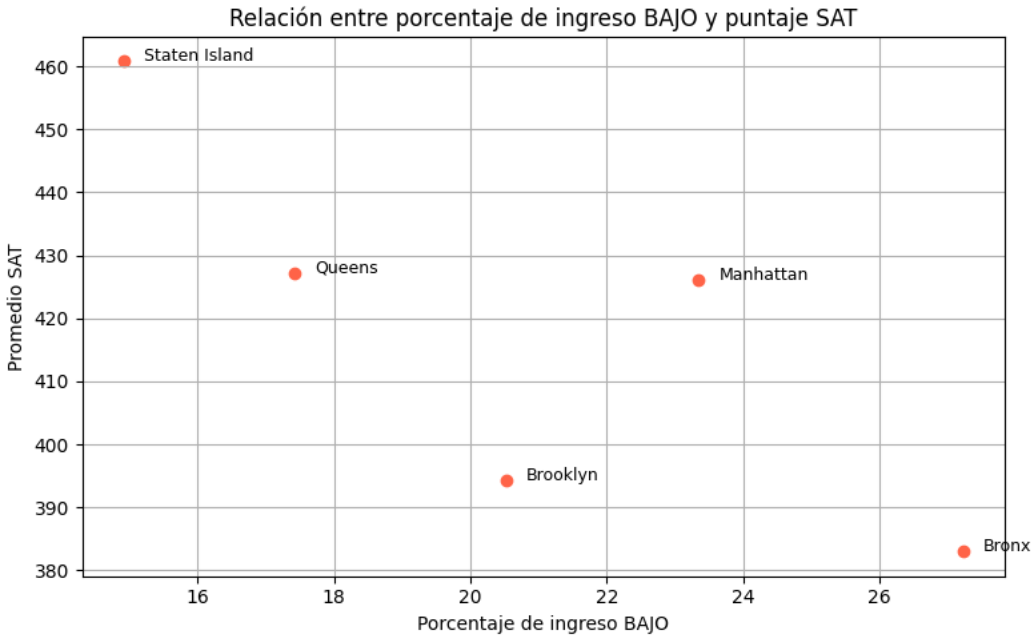

Figura 29: Relación entre porcentaje de ingreso bajo y puntaje SAT

La primera gráfica muestra una relación inversa entre el porcentaje de población con ingresos bajos y el puntaje promedio del SAT. A medida que aumenta la proporción de personas en situación de pobreza, el desempeño académico tiende a disminuir. Esto se aprecia claramente en el Bronx, donde el nivel de pobreza es el más alto y los puntajes son los más bajos, mientras que Staten Island se sitúa en el extremo opuesto.

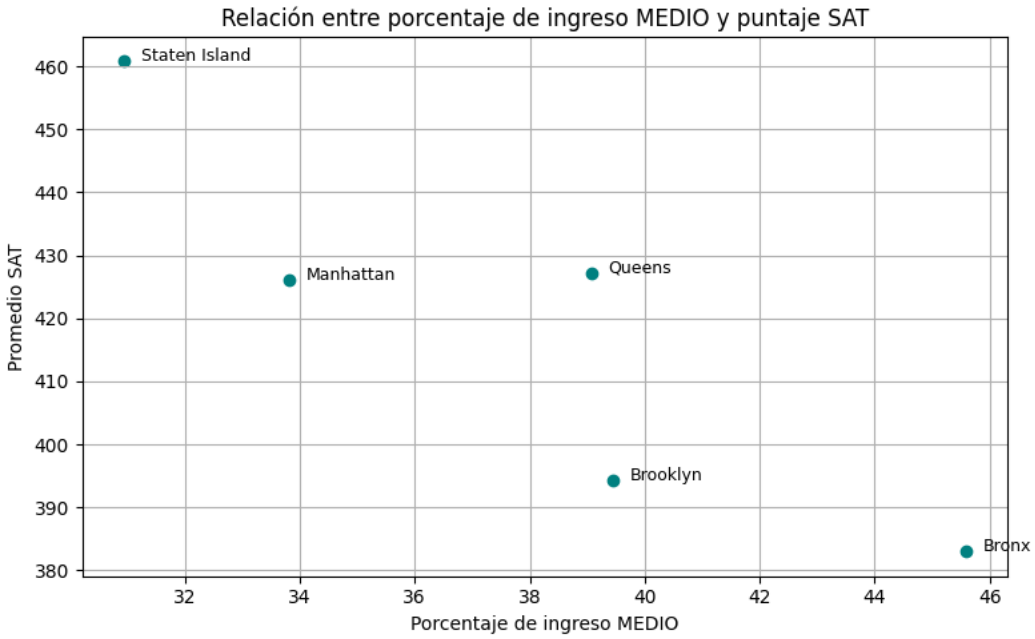

Figura 30: Relación entre porcentaje de ingreso medio y puntaje SAT

La segunda gráfica, que relaciona el porcentaje de ingresos medios con los puntajes del SAT, refuerza esta tendencia. Los distritos con una mayor proporción de población en niveles económicos medios o altos suelen obtener mejores resultados educativos. Sin embargo, la relación no es completamente lineal: Manhattan, por ejemplo, muestra un puntaje alto pese a no tener los valores más bajos en ingreso medio, lo que podría explicarse por la presencia de escuelas con mayores recursos o programas académicos especializados.

En conjunto, el análisis sugiere que el rendimiento académico está estrechamente vinculado con las condiciones económicas del entorno. Los distritos con menores niveles de pobreza y una base poblacional más estable tienden a ofrecer mejores oportunidades educativas, lo que refleja cómo las desigualdades socioeconómicas influyen de manera significativa en los resultados escolares de la ciudad.

14.6. Relación entre pobreza y arrestos por distrito

Para analizar si los distritos con mayores niveles de pobreza presentan también un número más alto de arrestos, se integraron dos fuentes de información: los datos del total de arrestos por distrito y los porcentajes de población clasificados por nivel de ingreso bajo y medio. Esta combinación permite observar de forma comparativa cómo se distribuye la actividad policial en relación con las condiciones económicas de cada zona.

La tabla resultante muestra diferencias marcadas entre los distritos. El Bronx y Brooklyn destacan con los porcentajes más altos de población en condición de

ingreso bajo (27,2 % y 20,5 %, respectivamente) y, al mismo tiempo, registran las cifras más altas de arrestos totales (31.882 y 39.086). En contraste, Staten Island presenta tanto el menor porcentaje de ingreso bajo (14,9 %) como la menor cantidad de arrestos (6.172). Los demás distritos, como Manhattan y Queens, se ubican en posiciones intermedias, lo que permite establecer una comparación más equilibrada entre contextos.

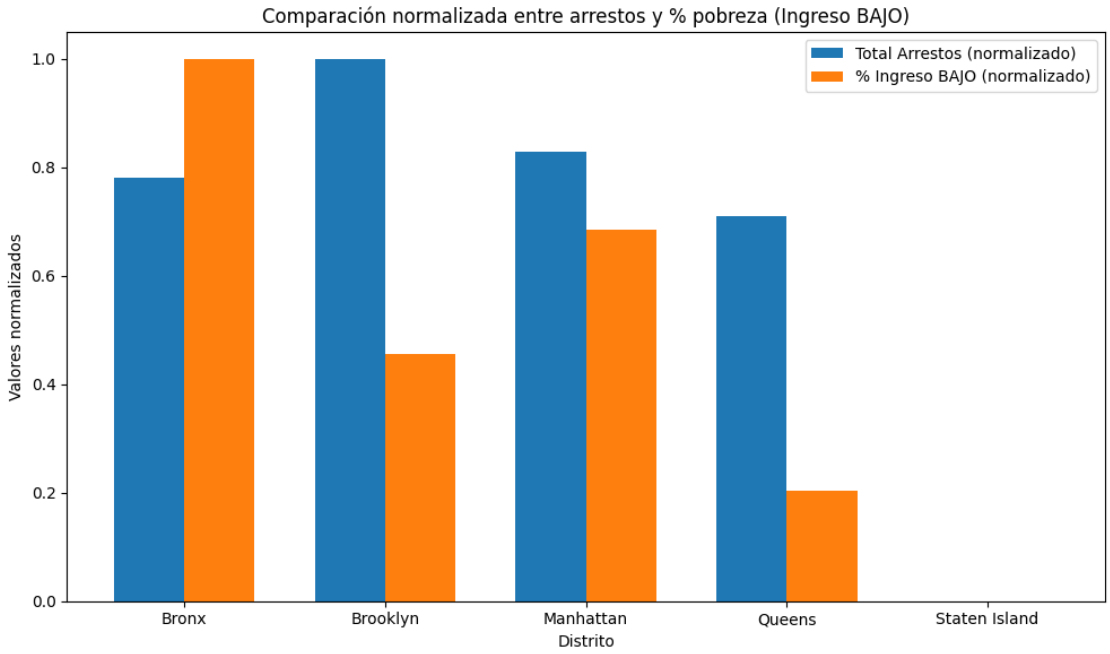

Figura 31: Comparación normalizada entre arrestos y porcentaje de pobreza (ingreso bajo)

La primera gráfica muestra los valores normalizados del total de arrestos y del porcentaje de población en ingreso bajo, facilitando la comparación entre distritos con tamaños poblacionales distintos. En ella se observa que el Bronx y Brooklyn destacan simultáneamente por sus altos niveles de pobreza y arrestos, mientras que Staten Island se mantiene en el extremo inferior de ambas variables. Queens y Manhattan presentan comportamientos intermedios, aunque con leves diferencias que sugieren la influencia de factores adicionales, como densidad urbana o intensidad de vigilancia.

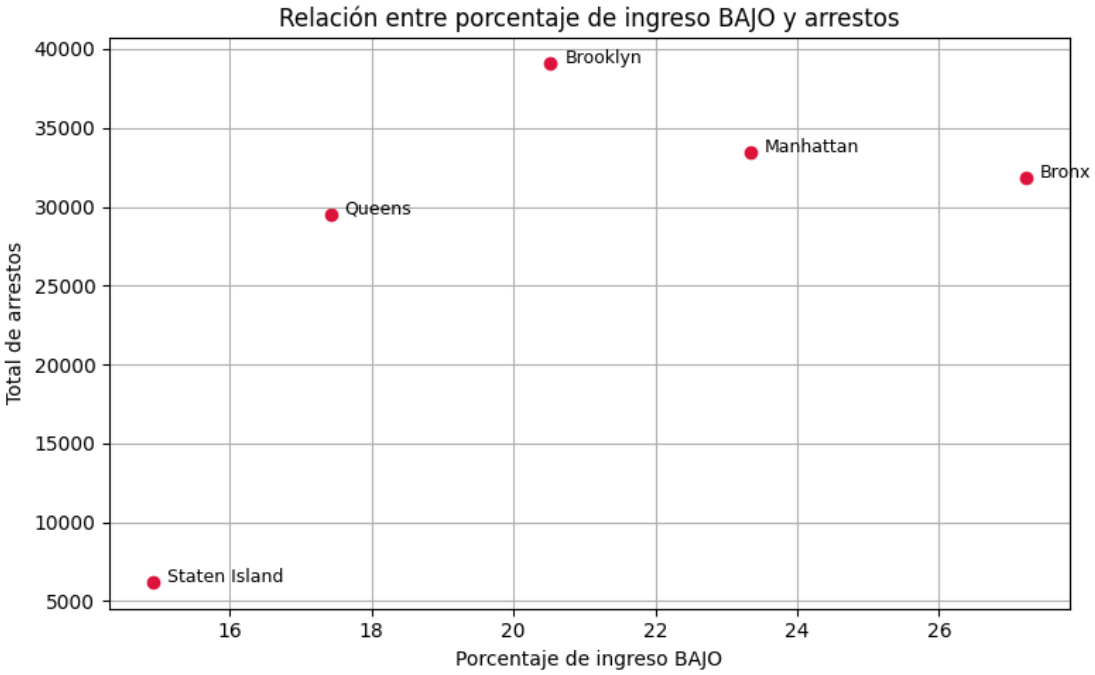

Figura 32: Relación entre porcentaje de ingreso bajo y total de arrestos

La gráfica de dispersión refuerza esta observación al mostrar una tendencia ascendente: a medida que aumenta el porcentaje de población en situación de pobreza, también lo hace el número total de arrestos. El Bronx, con el porcentaje más alto de ingreso bajo, se ubica entre los distritos con mayor número de arrestos, mientras que Staten Island, con los niveles de pobreza más bajos, presenta la menor incidencia.

En conjunto, los resultados sugieren una correlación positiva entre pobreza y arrestos: los distritos con condiciones económicas más desfavorables tienden a registrar un mayor número de detenciones. Sin embargo, esta relación no es completamente lineal, ya que factores como el tamaño de la población, la densidad urbana o las políticas locales de seguridad pueden influir en las diferencias observadas. Aun así, el patrón general evidencia cómo las desigualdades socioeconómicas pueden reflejarse también en la distribución de la actividad policial dentro de la ciudad.

14.7. Vehículos más involucrados en daños a propiedad pública

Con el objetivo de identificar qué tipos de vehículos participan con mayor frecuencia en accidentes que generan daños a la propiedad pública, se analizó el conjunto de datos de colisiones y se filtraron aquellos registros que reportaban este tipo de afectación. Posteriormente, los resultados se agruparon por categoría

de vehículo y se contó el número total de incidentes en cada una. Finalmente, se seleccionaron los quince tipos con mayor cantidad de registros, con el fin de representar de forma clara las tendencias principales.

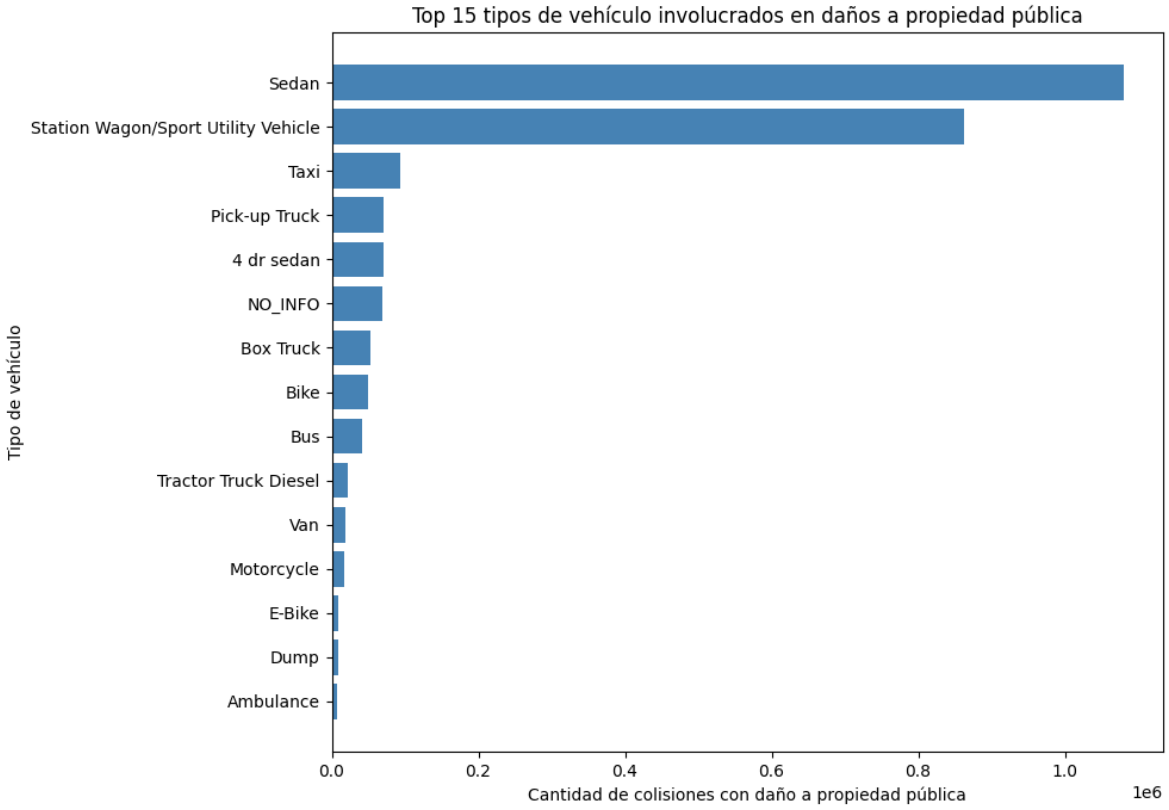

Figura 33: Top 15 tipos de vehículo involucrados en daños a propiedad pública

Los resultados muestran que los vehículos particulares, especialmente los *Sedan* y las camionetas tipo *Station Wagon/SUV*, encabezan la lista con una diferencia notable frente al resto. En conjunto, estos dos tipos superan ampliamente el millón y medio de incidentes, lo que refleja su predominio dentro del tráfico urbano de Nueva York. En tercer lugar aparecen los taxis, seguidos de camionetas tipo *Pick-up Truck* y sedanes de cuatro puertas, que aunque en menor cantidad, también representan una fracción significativa de los casos.

En posiciones intermedias se encuentran categorías como *Box Truck*, bicicletas, buses y camiones de carga diésel, lo cual indica que los vehículos de servicio o transporte pesado también contribuyen a los daños, aunque en menor proporción. Finalmente, los vehículos con menor participación incluyen motocicletas, bicicletas eléctricas, volquetas y ambulancias, lo que puede explicarse por su menor presencia en las vías o por un uso más controlado.

En general, la distribución evidencia que los automóviles particulares son los principales responsables de incidentes que afectan infraestructura o bienes públicos. Esto es coherente con su alta proporción dentro del parque automotor de la

ciudad, pero también sugiere la necesidad de fortalecer campañas de educación vial y estrategias de control dirigidas a conductores de este tipo de vehículos, especialmente en zonas residenciales o de alta densidad peatonal.

14.8. Análisis semanal de la ocurrencia de delitos violentos

Para identificar posibles patrones en la frecuencia de delitos violentos a lo largo de la semana, se agruparon los registros de arrestos según el día en que ocurrieron. En este análisis se consideraron únicamente los delitos clasificados como violentos, tales como asaltos, homicidios y violaciones, y se obtuvo el día correspondiente a partir de la fecha de arresto registrada en el conjunto de datos.

A partir de esta información, se construyó una tabla que resume el número total de arrestos por día, ordenada desde el lunes hasta el domingo. Esto permitió observar de manera clara cómo varía la incidencia de los delitos a lo largo de la semana.

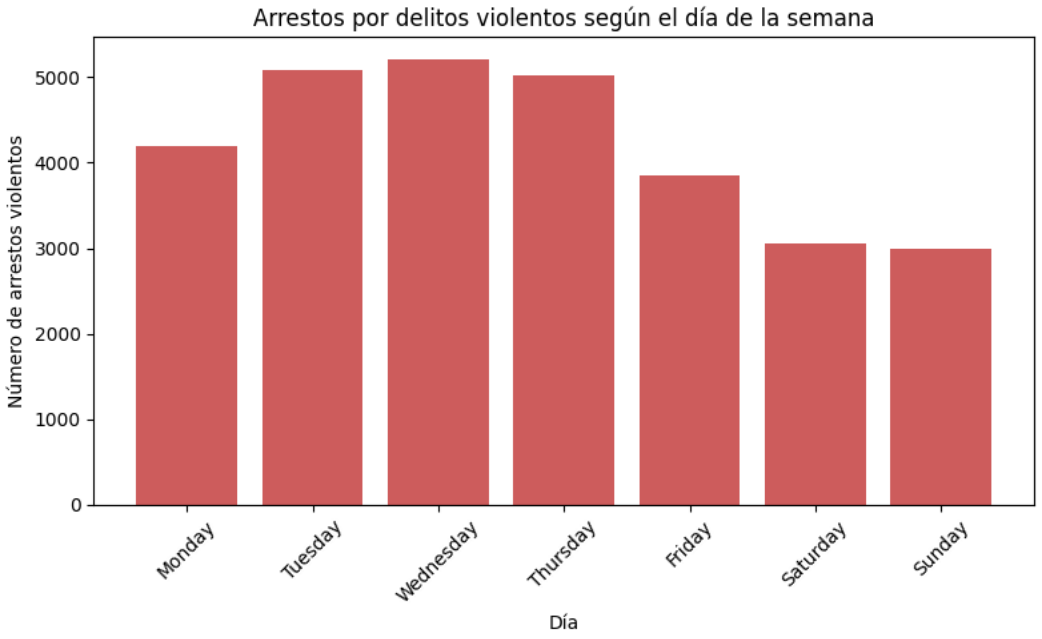

Figura 34: Arrestos por delitos violentos según el día de la semana

Los resultados muestran una tendencia marcada hacia una mayor cantidad de arrestos violentos durante los días hábiles, especialmente entre martes y jueves. El miércoles concentra el número más alto de casos, con 5.214 arrestos, seguido del martes (5.078) y el jueves (5.017). En contraste, los fines de semana presentan una reducción significativa en la actividad delictiva, con el sábado y el domingo como los días con menos registros (3.050 y 2.990, respectivamente).

Este patrón sugiere que los delitos violentos tienden a incrementarse en los días laborales, posiblemente asociados con el ritmo de la vida urbana, las interacciones sociales cotidianas o las dinámicas laborales que generan mayor exposición y conflictos. La disminución durante el fin de semana podría relacionarse con una menor densidad de tránsito peatonal y laboral en espacios públicos.

En conjunto, los resultados aportan información relevante para la gestión de la seguridad ciudadana, ya que permiten orientar los esfuerzos de vigilancia y prevención hacia los días de mayor riesgo, optimizando la distribución de recursos policiales según la tendencia observada.

15. Selección de Técnicas de Aprendizaje de Máquina

Con el propósito de complementar el análisis descriptivo realizado en las secciones anteriores, se incorporaron técnicas de aprendizaje de máquina que permitieran profundizar en la identificación de patrones y en la predicción de comportamientos dentro de los datos procesados. Para ello, se seleccionaron un algoritmo supervisado y uno no supervisado, aplicados sobre las variables más representativas de los conjuntos de datos.

15.1. Selección de algoritmos

Los algoritmos se escogieron considerando tanto la naturaleza de los datos como los objetivos del análisis:

- **Algoritmo supervisado — Regresión Logística:** se seleccionó por su capacidad para predecir la probabilidad de pertenecer al grupo de alta pobreza y, al mismo tiempo, ofrecer una interpretación clara del peso de cada variable.
- **Algoritmo no supervisado — K-Means:** se eligió porque permite identificar agrupamientos naturales en los datos, revelando perfiles socioeconómicos similares sin depender de etiquetas. Gracias a esto, es posible detectar patrones de desigualdad y concentraciones de pobreza dentro del territorio.

16. Preparación de Datos para Modelado

En esta etapa se finaliza el procesamiento de los datos con el fin de dejarlos listos para su uso en los modelos de aprendizaje automático.

16.1. Matriz de correlación y eliminación de variables redundantes

Para iniciar la preparación de los datos se seleccionaron las columnas numéricas relevantes del conjunto integrado, conformado por indicadores de pobreza, ingreso y variables socioeconómicas. En total se trabajó con 66 066 registros y las siguientes variables: `SEMP_adj`, `SSIP_adj`, `SSP_adj`, `WAGP_adj`, `NYCgov_Income`, `NYCgov_Threshold`, `Off_Threshold`, `PreTaxIncome_PU`, `TotalWorkHrs_PU` y `NYCgov_Income_Norm`.

Con estas columnas se construyó un vector de características mediante `VectorAssembler` y se calculó la matriz de correlación de Pearson utilizando la función `Correlation.corr` de Spark MLlib. El objetivo fue detectar relaciones lineales fuertes entre variables que pudieran introducir redundancia en los modelos.

El análisis evidenció correlaciones muy altas (—r— ¿0.9) entre las variables `NYCgov_Income`, `PreTaxIncome_PU` y `NYCgov_Income_Norm`, las cuales expresan información económica similar. De igual manera, `NYCgov_Threshold` y `Off_Threshold` mostraron una correlación elevada, al representar umbrales de ingreso estrechamente relacionados.

A partir de este resultado, se eliminaron las columnas `Off_Threshold`, `PreTaxIncome_PU` y `NYCgov_Income_Norm`, conservando aquellas que aportan información no redundante. Las variables finales seleccionadas para el modelado fueron: `SEMP_adj`, `SSIP_adj`, `SSP_adj`, `WAGP_adj`, `NYCgov_Income`, `NYCgov_Threshold` y `TotalWorkHrs_PU`.

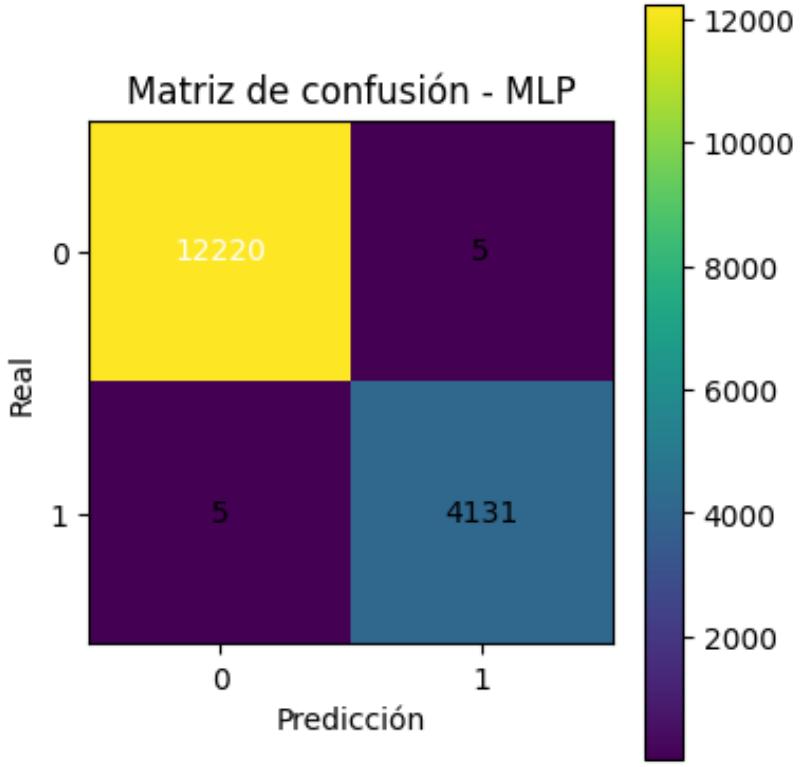

Figura 35: Matriz de correlación entre variables numéricas seleccionadas.

16.2. Normalización de variables numéricas

Con el subconjunto de variables finales se procedió a normalizar los datos para garantizar que todas las columnas tuvieran una contribución comparable en el proceso de modelado.

Primero, se ensamblaron las variables mediante un nuevo `VectorAssembler` en el vector `features_unnorm`, y luego se aplicó la clase `StandardScaler` de Spark MLlib. Este método centra las variables en media cero y las escala a varianza unitaria, evitando que aquellas con magnitudes mayores dominen el entrenamiento.

El resultado fue un nuevo conjunto de datos donde cada observación queda representada por un vector normalizado en la columna `features`, acompañado de la etiqueta de salida `label`. Este formato es el requerido por los modelos de aprendizaje supervisado y no supervisado a aplicar posteriormente.

Cuadro 3: Ejemplo de vector de características normalizado.

features	**label**
[-0.1208, -0.1630, -0.3541, 2.2194, 0.2537, -1.2608, -0.3500]	2

17. Aplicación de Modelos con MLlib

En esta sección se presentan los modelos de aprendizaje automático desarrollados con la biblioteca `MLlib` de Apache Spark, aplicados sobre los datos ya preparados y normalizados. El objetivo es explorar distintos enfoques de modelado (supervisados y no supervisados), los cuales permitan analizar la pobreza desde una perspectiva predictiva y estructural.

17.1. Modelo supervisado: implementación y resultados

Para la fase supervisada se construyó un modelo de **Regresión Logística** con el objetivo de clasificar los registros en dos grupos: *alta pobreza* (label = 1) y *baja pobreza* (label = 0). La variable objetivo se definió a partir del ingreso normalizado (`NYCgov_Income_Norm`), tomando como umbral el percentil 25 (0.2586). Aquellas observaciones con valores iguales o inferiores fueron etiquetadas como de alta pobreza, mientras que el resto fueron consideradas fuera de dicha condición.

El conjunto final estuvo compuesto por 66 066 registros, de los cuales 16 523 (25 %) correspondieron a la clase positiva y 49 543 (75 %) a la clase negativa, garantizando un balance razonable entre ambas categorías.

El conjunto de datos se dividió en 75 % para entrenamiento y 25 % para prueba, con el propósito de garantizar la validación del modelo. Se evaluaron tres métricas principales: Área Bajo la Curva ROC (AUC), Exactitud (Accuracy) y F1-score.

Los resultados iniciales mostraron un desempeño sobresaliente, con valores prácticamente perfectos (AUC = 1.0000, Accuracy = 0.9999 y F1-score = 0.9999). Si bien estas cifras indican una separación casi total entre las clases, también podrían sugerir cierto *overfitting*, aunque dentro del contexto del conjunto analizado se consideró aceptable.

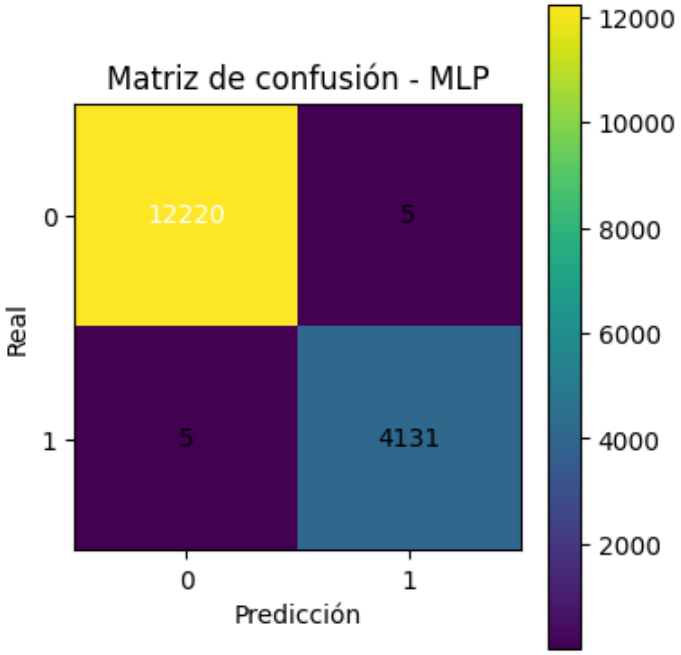

Figura 36: Matriz de confusión del modelo de Regresión Logística.

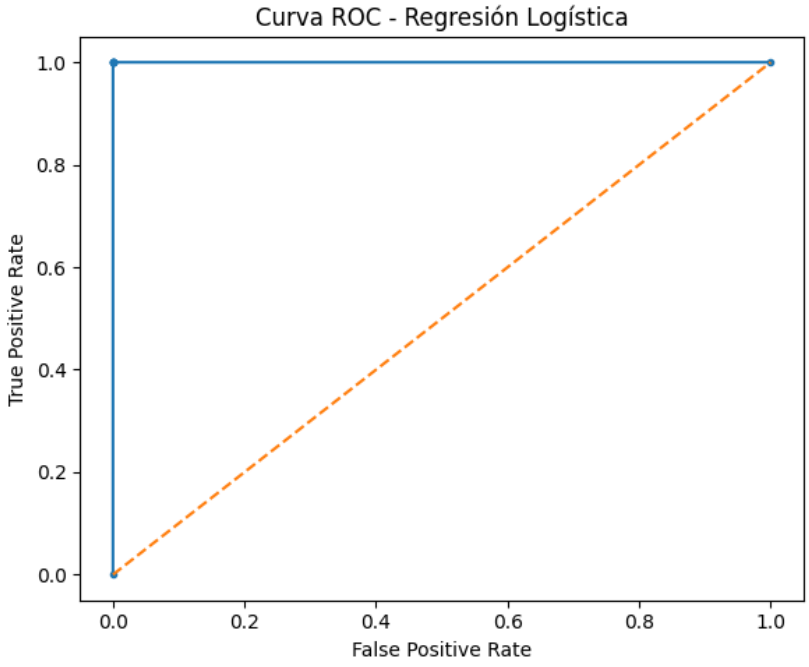

Figura 37: Curva ROC de la Regresión Logística.

El modelo asignó coeficientes altos a variables asociadas directamente con el ingreso, destacándose `NYCgov_Income` como el predictor más influyente, con un peso negativo considerable (-45 836.59). Esto implica que, a menor ingreso, mayor es la probabilidad de pertenecer al grupo clasificado como de alta pobreza.

Las demás variables presentaron contribuciones menores pero coherentes con su relación económica:

Cuadro 4: Coeficientes del modelo de Regresión Logística.

Variable	**Coeficiente**
NYCgov_Income	-45 836.59
SEMP_adj	29.98
NYCgov_Threshold	10.87
SSIP_adj	9.64
SSP_adj	7.26
TotalWorkHrs_PU	2.91
WAGP_adj	-0.25

Posteriormente, se exploraron diferentes combinaciones de hiperparámetros para analizar la sensibilidad del modelo ante la regularización. Los parámetros ajustados fueron:

- `regParam`: intensidad de regularización (valores 0.000, 0.010 y 0.100),
- `elasticNetParam`: mezcla entre regularización L1 y L2 (valores 0.0, 0.5 y 1.0).

Los resultados se mantuvieron estables a lo largo de todas las configuraciones, mostrando valores de AUC entre 0.96 y 1.00 y F1-score superiores a 0.85 en todos los casos. Esto evidencia que el modelo conserva una gran capacidad de discriminación entre clases incluso bajo distintas condiciones de penalización.

Cuadro 5: Resultados de la búsqueda de hiperparámetros.

regParam	**elasticNetParam**	**AUC**	**F1-score**
0.000	0.0	1.0000	1.0000
0.010	0.0	0.9868	0.9440
0.100	0.0	0.9618	0.8817
0.010	0.5	0.9930	0.9586
0.100	0.5	0.9668	0.8667
0.010	1.0	0.9998	0.9910
0.100	1.0	0.9894	0.8539

17.2. Modelo no supervisado: K-Means

Para complementar el análisis predictivo, se aplicó un modelo no supervisado basado en el algoritmo **K-Means**, con el propósito de identificar grupos o patrones dentro de los datos sin utilizar la variable de pobreza como referencia.

El modelo se entrenó sobre el vector de características normalizado obtenido en etapas anteriores, probando diferentes valores de K (número de grupos) entre 2 y

8. Para cada configuración se calculó la métrica de *silhouette*, que mide el grado de separación y cohesión interna de los clusters. Valores cercanos a 1 indican agrupaciones bien diferenciadas, mientras que valores bajos o negativos reflejan solapamiento entre grupos.

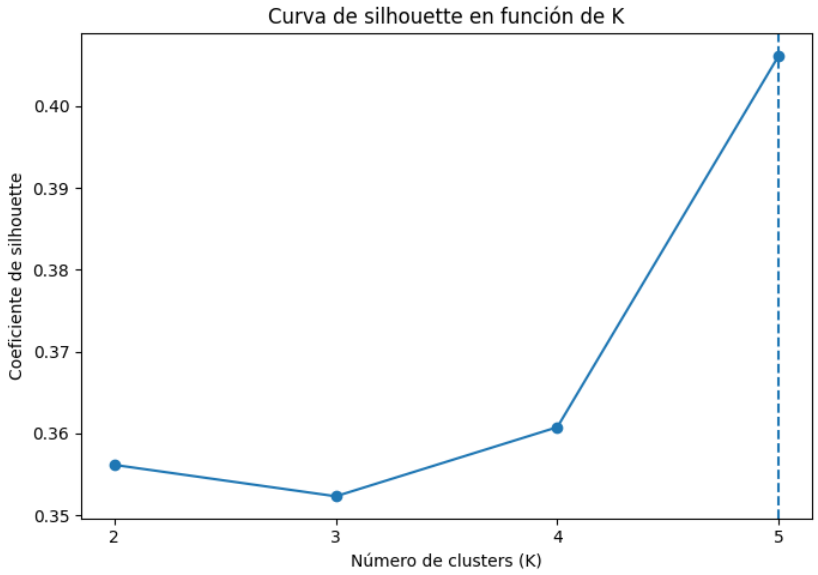

Figura 38: Evolución del índice silhouette en función del número de clusters.

Como se observa en la Figura 38, el mejor desempeño se obtuvo con $K = 5$, alcanzando un valor de silhouette de 0.4062. Este resultado indica una separación moderadamente clara entre los grupos, lo que sugiere la existencia de distintos perfiles económicos dentro de la población analizada.

Cuadro 6: Evaluación del modelo K-Means con diferentes valores de K.

Número de clusters (K)	**Índice silhouette**
2	0.3562
3	0.3523
4	0.3608
5	**0.4062**

Los centros de los cinco grupos (en el espacio normalizado) muestran diferencias notables en variables como ingresos, salarios y horas de trabajo, evidenciando la diversidad de perfiles socioeconómicos presentes en la ciudad.

Cuadro 7: Centros de los clusters (variables normalizadas).

Cluster	Centro del cluster (features normalizadas)
0	[-0.0156, -0.1431, -0.2914, -0.2620, -0.7217, -0.4607, 0.6051]
1	[-0.0350, -0.1621, -0.3305, 1.9060, 0.9632, -0.4117, -0.6405]
2	[0.0522, -0.1527, -0.2894, -0.2947, 0.4716, 0.8672, -0.7931]
3	[-0.0538, -0.1439, 2.6761, -0.5042, -0.1867, -0.7019, 1.2177]
4	[-0.1035, 5.5684, 0.0063, -0.5806, -0.5574, -0.4214, 1.0772]

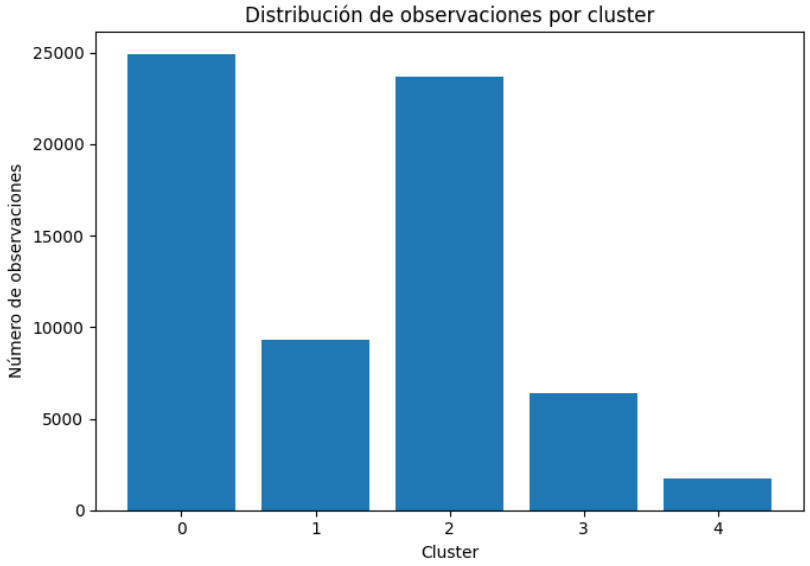

Figura 39: Distribución de observaciones por cluster en el modelo K-Means.

La Figura 39 muestra el tamaño de cada grupo. El `Cluster 0` fue el más numeroso, con 24 913 registros, seguido por los clusters `2` y `1`, mientras que los clusters `3` y `4` concentraron los casos menos frecuentes.

Para interpretar los resultados, se calculó la **tasa promedio de pobreza aproximada** dentro de cada grupo, combinando la asignación de cada registro a su cluster con la variable `label`. Esta medida permite identificar qué tan concentrados están los casos de alta pobreza en cada segmento.

Cuadro 8: Promedio de pobreza por cluster.

Cluster	Número de observaciones	Tasa de pobreza aproximada
0	24 913	1.64
1	9 305	2.00
2	23 697	1.92
3	6 424	1.87
4	1 727	1.64

Los resultados muestran diferencias claras entre los cinco grupos. Los clusters `1` y `2` presentan mayores valores promedio, lo que indica una concentración más alta de registros asociados a bajos ingresos o condiciones de vulnerabilidad. En contraste, los clusters `0` y `4` reflejan perfiles con mejores indicadores económicos, mientras que el cluster `3` se ubica en una posición intermedia.

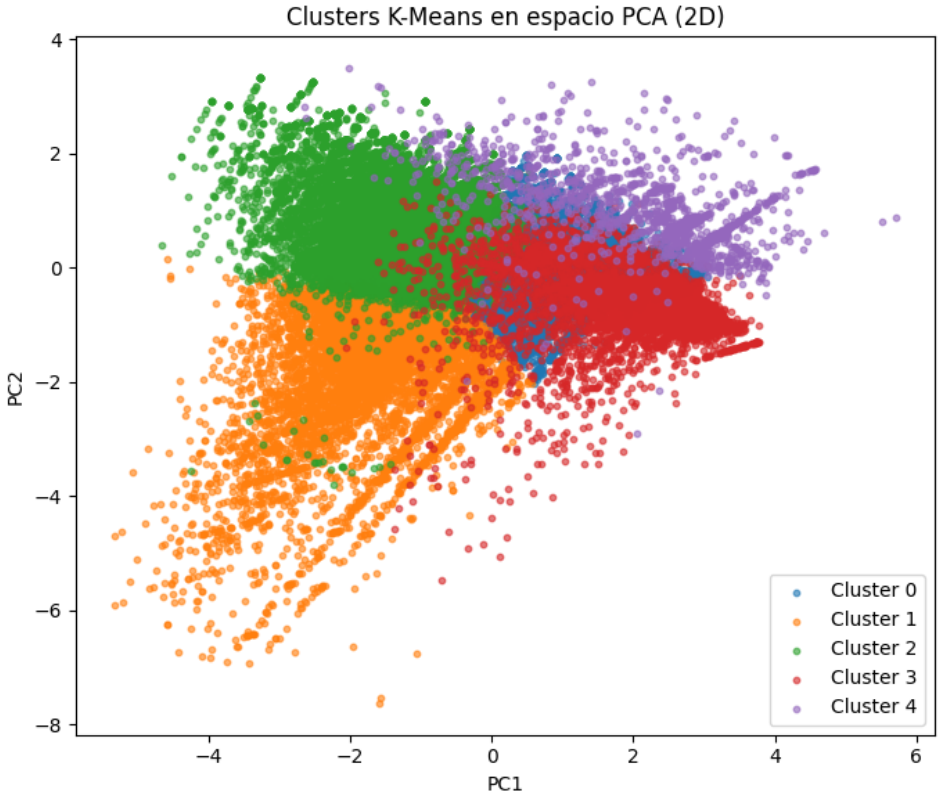

Figura 40: Visualización de los clusters K-Means en espacio PCA (2D).

Finalmente, la Figura 40 muestra la representación bidimensional de los cinco clusters proyectados mediante Análisis de Componentes Principales (PCA). La separación visual observada entre los grupos confirma los resultados numéricos obtenidos con el índice *silhouette*, destacando la presencia de patrones socioeconómicos diferenciados dentro de la población.

18. Implementación de Deep Learning

Tras aplicar los modelos supervisado y no supervisado, se implementó un modelo de **aprendizaje profundo** con el fin de explorar si una red neuronal podía mejorar la capacidad de clasificación y capturar relaciones no lineales entre las variables socioeconómicas. El objetivo principal del modelo fue predecir la probabilidad de que una observación correspondiera a un hogar en condición de *alta pobreza*, a partir de los indicadores numéricos previamente normalizados.

18.1. Arquitectura y configuración

El modelo desarrollado consistió en una red neuronal con varias capas ocultas entrenadas sobre el conjunto de datos procesado en `Spark MLlib`. Las variables de entrada fueron las siguientes: `SEMP_adj`, `SSIP_adj`, `SSP_adj`, `WAGP_adj`, `NYCgov_Income`, `NYCgov_Threshold` y `TotalWorkHrs_PU`.

La arquitectura general incluyó:

- Capas ocultas densas con activación `ReLU`.
- Capa de salida con activación `sigmoid` para clasificación binaria.
- Optimizador `Adam` y función de pérdida `binary_crossentropy`.
- Regularización mediante *dropout* y normalización por lotes.

18.2. Preparación de datos de entrada

Previo al entrenamiento, los datos se dividieron en conjuntos de entrenamiento y prueba (70 %-30 %), manteniendo la proporción entre las clases `label = 0` (no alta pobreza) y `label = 1` (alta pobreza). Todas las variables numéricas se estandarizaron con `StandardScaler`, y se excluyó la columna `NYCgov_Income_Norm` para evitar redundancias con otras variables de ingreso.

La Figura 41 muestra la distribución de las observaciones en un espacio reducido a dos componentes principales mediante PCA, donde se aprecia la separación aproximada entre hogares clasificados como de alta y no alta pobreza.

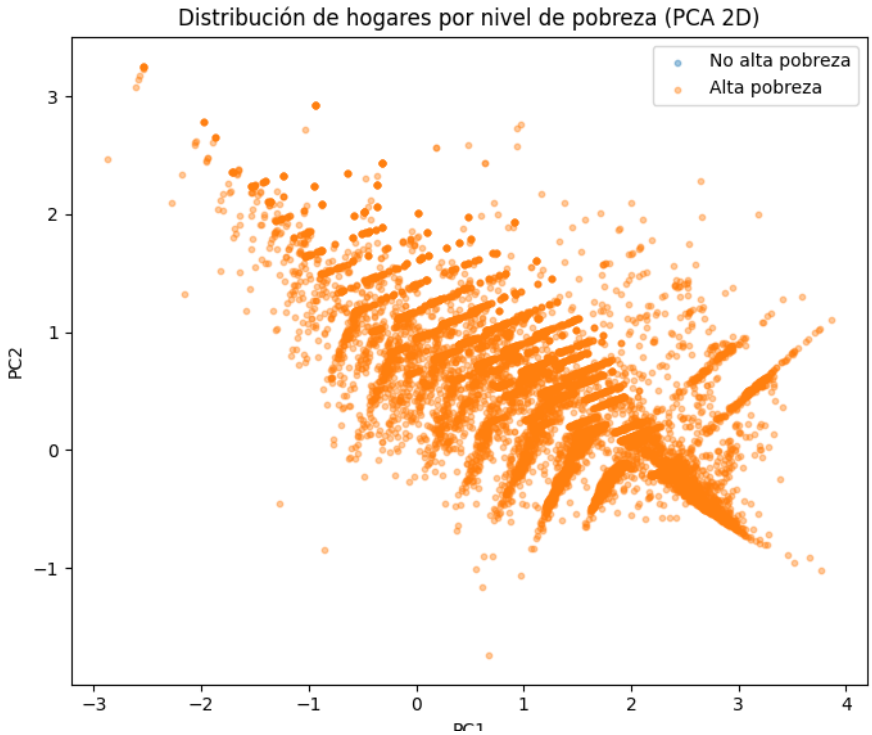

Figura 41: Distribución de hogares por nivel de pobreza proyectados en PCA (2D).

18.3. Resultados y comparación

El modelo logró un desempeño sobresaliente, con métricas cercanas a la perfección:

Cuadro 9: Métricas de desempeño del modelo MLP.

Métrica	**Valor**
AUC ROC	1.0000
Accuracy	0.9994
F1-score	0.9994

Estos valores evidencian que el MLP clasificó correctamente casi todos los registros del conjunto de prueba, replicando el rendimiento alcanzado por la regresión logística.

19. Evaluación de Modelos

La evaluación final de los modelos aplicados busca comparar su desempeño y pertinencia frente al objetivo general del proyecto: identificar patrones y factores asociados a la pobreza en la ciudad de Nueva York. Para ello, se consideraron tres enfoques complementarios:

- Un modelo **supervisado lineal** (Regresión Logística), enfocado en la interpretación y explicación de las variables socioeconómicas.
- Un modelo **de aprendizaje profundo** (Perceptrón Multicapa, MLP), capaz de capturar relaciones no lineales y complejas.
- Un modelo **no supervisado** (K-Means), orientado a descubrir agrupamientos naturales dentro de los datos sin utilizar etiquetas.

Cada modelo fue evaluado según las métricas más relevantes para su tipo: AUC, exactitud y F1-score para los modelos supervisados, e índice de *silhouette* para el modelo no supervisado. La Tabla 10 presenta un resumen general de los resultados obtenidos.

Cuadro 10: Comparación general de desempeño entre modelos aplicados.

Modelo	**Tipo**	**AUC ROC**	**Accuracy / Silhouette**	**F1-score**
Regresión Logística	Supervisado	1.0000	0.9999	0.9999
Red Neuronal (MLP)	Supervisado (profundo)	1.0000	0.9994	0.9994
K-Means (K=5)	No supervisado	—	0.4062 (Silhouette)	—

Los resultados reflejan un desempeño excepcional de los modelos supervisados, con métricas casi perfectas en todos los casos. Tanto la Regresión Logística como la Red Neuronal logran diferenciar de forma precisa entre hogares de alta y no alta pobreza.

Por otro lado, el modelo no supervisado K-Means (sin etiquetas) logró segmentar los datos en grupos con diferencias socioeconómicas claras, alcanzando un valor de *silhouette* de 0.4062.

En conjunto, la combinación de modelos supervisados, no supervisados y de aprendizaje profundo permitió abordar el problema desde múltiples perspectivas: predicción, interpretación y descubrimiento de patrones, aportando una visión más completa de la dinámica socioeconómica en la ciudad.

20. Conclusiones y Hallazgos encontrados

El análisis de los distintos conjuntos de datos permitió identificar patrones consistentes entre las condiciones socioeconómicas, educativas y de seguridad de la ciudad de Nueva York. Se observó que la desigualdad territorial sigue siendo un factor determinante en la distribución de los problemas urbanos: los distritos con mayores niveles de pobreza, como el Bronx y Brooklyn, presentan también un rendimiento educativo inferior y una mayor incidencia de delitos. En contraste, áreas como Manhattan y Staten Island concentran mejores ingresos, puntajes académicos más altos y una menor tasa de arrestos, reflejando una clara segmentación entre zonas de alta y baja oportunidad social.

En el caso de la educación, los resultados mostraron diferencias visibles entre distritos, aunque no tan marcadas como las asociadas a la pobreza o la seguridad. Las instituciones con menores puntajes promedio en el examen SAT se ubican principalmente en sectores con menor ingreso y mayores problemas sociales, lo que sugiere que más allá del sistema educativo, el entorno en el que se desarrollan los estudiantes influye de manera importante en su desempeño. También se identificaron zonas donde los resultados académicos son mejores de lo esperado pese a las limitaciones del contexto, lo que evidencia la influencia de factores locales como el apoyo familiar o comunitario.

El estudio de los accidentes viales mostró que la mayoría de los eventos se relacionan principalmente con el comportamiento humano, en especial por causas como la distracción o el desconocimiento de las normas de tránsito. Además, la concentración de incidentes en días laborales y horarios de alta movilidad sugiere que la densidad vehicular y la presión del tráfico diario influyen de manera significativa en la ocurrencia de estos sucesos.

Desde la perspectiva analítica, los modelos aplicados revelaron una fuerte relación entre las variables socioeconómicas y las condiciones de vulnerabilidad urbana. La capacidad predictiva obtenida a partir de los datos respalda el uso de técnicas de aprendizaje automático como herramienta para interpretar dinámicas sociales complejas y anticipar escenarios de riesgo. En conjunto, los hallazgos ponen de relieve la interdependencia entre pobreza, educación, seguridad y movilidad, destacando la necesidad de estrategias integradas que aborden estos factores de manera coordinada en el contexto urbano de Nueva York.

21. Recomendaciones

A partir de los resultados obtenidos, se propone orientar las acciones hacia la reducción de las desigualdades territoriales que atraviesan la ciudad de Nueva York. Los distritos con mayores niveles de pobreza y menor acceso a oportunidades necesitan políticas públicas que combinen inversión social con mejoras en educación, seguridad y movilidad. En lugar de tratar cada tema por separado, las intervenciones deberían centrarse en fortalecer las comunidades y los entornos donde estos problemas coinciden, buscando soluciones que integren lo social, lo urbano y lo educativo.

En el plano social, se sugiere impulsar programas que fortalezcan las redes locales de apoyo, generen empleo y promuevan la participación ciudadana. Estas iniciativas pueden contribuir a disminuir la vulnerabilidad de los barrios más afectados por la pobreza y la violencia. También sería conveniente aumentar la presencia institucional en estos sectores mediante proyectos de acompañamiento familiar, oferta cultural y presencia policial enfocada en la prevención y la convivencia.

En el ámbito educativo, se recomienda mejorar las condiciones de las escuelas ubicadas en zonas con menores ingresos, tanto en infraestructura como en re-

cursos pedagógicos. Si bien la educación no aparece como la causa principal de las diferencias observadas, sí refleja el impacto del entorno en el que viven los estudiantes.

En cuanto a la movilidad urbana, los resultados muestran que la mayoría de los accidentes se originan por errores humanos, especialmente por distracciones, exceso de confianza o conductas imprudentes al volante. Este tipo de comportamientos requiere una respuesta que combine una mejor educación vial desde edades tempranas con una aplicación más estricta de las normas de tránsito. Programas de formación continua, sanciones efectivas y campañas que refuercen la responsabilidad de peatones y conductores pueden ayudar a modificar hábitos de riesgo. Además, sería útil reforzar la supervisión en los horarios de mayor congestión, especialmente durante los días laborales, cuando el tráfico intenso y el estrés diario aumentan el riesgo de siniestros. Mejorar la señalización, la iluminación y el diseño de las vías también puede reducir estos eventos, pero el cambio más profundo depende de promover una cultura vial basada en la responsabilidad compartida.

Finalmente, es necesario fortalecer la gestión y calidad de los datos públicos que alimentan los análisis sociales y urbanos. Una base de datos más completa y estandarizada permitiría tomar decisiones más acertadas y medir con precisión el impacto de las políticas implementadas. Se recomienda registrar de forma consistente las zonas geográficas en los reportes de accidentes, reducir los valores nulos mediante sistemas de validación, y mantener un control sobre la actualización y veracidad de la información. Contar con datos limpios y bien estructurados garantiza que los esfuerzos analíticos futuros reflejen mejor la realidad de la ciudad y sirvan como guía para intervenciones más efectivas.

Referencias

- Datos oficiales: https://data.cityofnewyork.us
- CRISP-DM: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=SS3RA7_Cloud/com.ibm.spss.modeler.help/idh_crispdm_main.htm
- OpenWeatherMap API: https://openweathermap.org/api
- Wikipedia - Nueva York: https://es.wikipedia.org/wiki/Nueva_York
- Blog SparklyMaid NYC: https://www.sparklymaidnyc.com/blog/que-lugar-ocupa-nu
- Información general: https://www.nuevayork.net/informacion-general
- Seguridad vial en NY: https://www.semana.com/mundo/noticias-estados-unidos/articulo/nueva-york-lidera-la-seguridad-vial-en-estados-unidos-logro-conve202517/
- Crisis de personas sin techo: https://www.france24.com/es/programas/enlace/20240906-desolaci%C3%B3n-y-desamparo-la-crisis-de-los-sintecho-en-n
- Desigualdad en NY: https://www.bbc.com/mundo/noticias/2016/04/160418_nueva_york_ricos_pobres_primarias_ps
- Riesgos sociales en NY: https://nychazardmitigation.com/es/documentation/nyc-hazard-environment/social/
- Tasa de desempleo (EE.UU.): https://es.tradingeconomics.com/united-states/unemployment-rate
- Estadísticas de NY: https://datacommons.org/place/geoId/36?hl=es
- Estimaciones de población NYC 2025: https://www.nyc.gov/assets/planning/downloads/pdf/our-work/reports/new-york-city-population-estimates-and-tren may-2025.pdf