Proyecto Robos Y Accidentes De Tráfico En La Ciudad De Nueva York

Juan Jose Rodriguez

Daniel Andrés Becerra Sierra

Jhonatan Rodrigo Robayo Barrera

Miguel Mendez Hernandez

Pontificia Universidad Javeriana

Procesamiento de Datos a Gran Escala

Ciencia de Datos

13/03/2024

Entendimiento del negocio

Nueva York es uno de los estados más poblados y económicamente importantes de los Estados Unidos. Con una población diversa y una economía que abarca desde la industria financiera hasta la tecnología y el turismo, enfrenta desafíos únicos en cuanto a la gestión urbana y territorial. Los centros urbanos densamente poblados, como la Ciudad de Nueva York, enfrentan desafíos específicos relacionados con el crimen, la seguridad vial y la infraestructura. Posee una población de aproximadamente 8.468 millones de habitantes.

Datos Macroeconomicos

- **Crecimiento del Empleo**: Aunque Nueva York experimentó un aumento significativo en la creación de empleo en 2022, con más de 443,000 trabajos añadidos (un aumento del 5%), aún no ha logrado recuperar los niveles de empleo pre-pandemia, quedando un 3% por debajo de los niveles de 2019.
- **Tasa de Desempleo**: En noviembre de 2023, la tasa de desempleo en Nueva York fue del 4.3%, superior al promedio nacional, indicando un mercado laboral aún en recuperación.
- **Crecimiento del PIB**: Para el tercer trimestre de 2023, el crecimiento real del PIB de Nueva York fue del 3.5% anual, mostrando un crecimiento económico más bajo que el promedio nacional. El PIB real de Nueva York alcanzó los $2.0 trillones en bienes y servicios por año.
- **Crecimiento de Ingresos Personales**: Nueva York se clasificó en el puesto 45 a nivel nacional en términos de crecimiento de ingresos personales en 2022, con un crecimiento del 0.8%, lo cual se atribuye en parte a la finalización de los pagos de impacto económico y los beneficios de desempleo mejorados en 2021.
- **Disminución de la Población**: En 2022, la población de Nueva York continuó su tendencia a la baja, disminuyendo en un 0.9% desde 2021, contrastando con un aumento del 0.4% a nivel nacional.

Objetivo

Identificar diferentes factores detonantes o relacionales específicos de accidentes y los principales casos de delitos además de las relaciones potenciales entre diferentes datos demográficos y de pobreza con estos eventos, para así poder desarrollar un plan de acción detallado y basado en análisis de datos que permita al estado de Nueva York abordar estos problemas de manera efectiva y permitiendo la reducción de los mismos.

Selección de los datos a utilizar

NYPD Arrest Data (Year to Date)

Este conjunto de datos proporciona información detallada sobre cada arresto realizado por el NYPD durante el año en curso, incluyendo el tipo de crimen, la ubicación, la hora del arresto, y datos demográficos del sospechoso. Es esencial para identificar patrones y tendencias en los arrestos, lo que permite entender mejor cómo y dónde se concentran los delitos. Además, facilitará la identificación de posibles correlaciones entre arrestos y factores socioeconómicos, contribuyendo al desarrollo de estrategias de prevención específicas.

Motor Vehicle Collisions - Vehicles

Este conjunto de datos contiene detalles sobre cada vehículo involucrado en colisiones desde abril de 2016. Cada registro representa un vehículo involucrado en un accidente, ofreciendo una visión profunda sobre las causas y circunstancias de los accidentes viales. La inclusión de este conjunto de datos es fundamental para analizar las tendencias y patrones de accidentes viales, identificar factores de riesgo comunes.

Colección y descripción de datos

NYPD Arrest Data (Year to Date)

- ARREST_KEY: integer (nullable = true). Identificador único y persistente generado aleatoriamente para cada arresto.
- ARREST_DATE: date (nullable = true) Fecha exacta del arresto.
- PD_CD: integer (nullable = true) Código de clasificación interna de tres dígitos, más detallado que el código KY. Ayuda a identificar el tipo específico de delito por el cual se efectuó el arresto.
- PD_DESC: string (nullable = true) Descripción de la clasificación interna correspondiente al código PD, proporcionando detalles más granulares sobre el delito.
- KY_CD: integer (nullable = true) Código de clasificación interna de tres dígitos que indica una categoría más general de delito que el código PD.
- OFNS_DESC: string (nullable = true) Descripción de la clasificación interna correspondiente al código KY, indicando la categoría más general del delito
- LAW_CODE: string (nullable = true) Códigos de la ley bajo los cuales se efectuaron los cargos, incluyendo la Ley Penal de NYS, VTL (Ley de Tránsito Vehicular) y otras leyes locales.
- LAW_CAT_CD: string (nullable = true) Nivel del delito (felonía, delito menor, infracción)
- ARREST_BORO: string (nullable = true) Abreviatura del borough de Nueva York donde se realizó el arresto (Bronx, Staten Island, Brooklyn, Manhattan, Queens).
- ARREST_PRECINCT: integer (nullable = true) Precinto donde ocurrió el arresto
- JURISDICTION_CODE: integer (nullable = true) Código de la jurisdicción responsable del arresto, donde los códigos 0 (Patrulla), 1 (Tránsito) y 2 (Vivienda) representan jurisdicciones del NYPD, y códigos 3 en adelante representan otras jurisdicciones.
- AGE_GROUP: string (nullable = true) Grupo de edad del perpetrador

- PERP_SEX: string (nullable = true) Grupo de edad del perpetrador
- PERP_RACE: string (nullable = true) Raza del perpetrador
- X_COORD_CD: integer (nullable = true) Coordenadas de la ubicación del arresto en el Sistema de Coordenadas Planas del Estado de Nueva York
- Y_COORD_CD: integer (nullable = true) Coordenadas de la ubicación del arresto en el Sistema de Coordenadas Planas del Estado de Nueva York
- Latitude: double (nullable = true) Coordenadas de latitud para el sistema de coordenadas global
- Longitude: double (nullable = true)Coordenadas de longitud para el sistema de coordenadas global
- New Georeferenced Column: string (nullable = true) Columna que combina la latitud y longitud en un solo punto georreferenciado

El dataset cubre desde la identificación precisa de cada arresto hasta detalles complejos sobre delitos, leyes aplicadas, y severidad de las ofensas. Incluye información demográfica detallada de los perpetradores, como edad, sexo, y raza, permitiendo análisis profundos sobre patrones criminales entre diferentes grupos. Además, proporciona datos geográficos precisos, desde la ubicación específica del arresto hasta el distrito y el precinto, facilitando el estudio de la distribución espacial de los crímenes y las operaciones policiales.

Motor Vehicle Collisions - Vehicles

- UNIQUE_ID: integer (nullable = true). Código único de registro generado por el sistema, que sirve como clave primaria.
- COLLISION_ID: integer (nullable = true). Código de identificación del choque. Clave foránea que coincide con unique_id de la tabla de choques.
- CRASH_DATE: date (nullable = true). Fecha de ocurrencia de la colisión.
- CRASH_TIME: timestamp (nullable = true). Hora de ocurrencia de la colisión.
- VEHICLE_ID: string (nullable = true). Código de identificación del vehículo asignado por el sistema.
- STATE_REGISTRATION: string (nullable = true). Estado donde el vehículo está registrado.
- VEHICLE_TYPE: string (nullable = true). Tipo de vehículo basado en la categoría de vehículo seleccionada (ATV, bicicleta, carro/suv, ebike, escooter, camión/bus, motocicleta, otro).
- VEHICLE_MAKE: string (nullable = true). Marca del vehículo.
- VEHICLE_MODEL: string (nullable = true). Modelo del vehículo.
- VEHICLE_YEAR: integer (nullable = true). Año de fabricación del vehículo.
- TRAVEL_DIRECTION: string (nullable = true). Dirección hacia la cual viajaba el vehículo.
- VEHICLE_OCCUPANTS: integer (nullable = true). Número de ocupantes en el vehículo.
- DRIVER_SEX: string (nullable = true). Sexo del conductor.
- DRIVER_LICENSE_STATUS: string (nullable = true). Estado de la licencia del conductor (licencia, permiso, no licenciado).
- DRIVER_LICENSE_JURISDICTION: string (nullable = true). Estado donde fue emitida la licencia del conductor.
- PRE_CRASH: string (nullable = true). Acción previa al choque (ir recto, girar a la derecha, adelantar, retroceder, etc.).

- POINT_OF_IMPACT: string (nullable = true). Ubicación en el vehículo del punto inicial de impacto (por ejemplo, lado del conductor, lado trasero del pasajero, etc.).
- VEHICLE_DAMAGE: string (nullable = true). Ubicación en el vehículo donde ocurrió la mayor parte del daño.
- VEHICLE_DAMAGE_1: string (nullable = true). Ubicaciones adicionales del daño en el vehículo.
- VEHICLE_DAMAGE_2: string (nullable = true). Ubicaciones adicionales del daño en el vehículo.
- VEHICLE_DAMAGE_3: string (nullable = true). Ubicaciones adicionales del daño en el vehículo.
- PUBLIC_PROPERTY_DAMAGE: string (nullable = true). Daño a propiedad pública (Sí o No).
- PUBLIC_PROPERTY_DAMAGE_TYPE: string (nullable = true). Tipo de propiedad pública dañada (por ejemplo, señal, cerca, poste de luz, etc.).
- CONTRIBUTING_FACTOR_1: string (nullable = true). Factores que contribuyen a la colisión para el vehículo designado.
- CONTRIBUTING_FACTOR_2: string (nullable = true). Factores adicionales que contribuyen a la colisión para el vehículo designado.

Este dataset proporciona información detallada sobre incidentes de tráfico en una ubicación geográfica específica, recopilando datos de cada vehículo involucrado en colisiones. Este dataset incluye identificadores únicos para registros y colisiones, fechas y horas de los incidentes, así como información específica del vehículo como tipo, marca, modelo, año, y dirección de viaje. Además, se recoge información sobre los ocupantes, el sexo del conductor, el estado de su licencia y la jurisdicción de emisión, acciones previas al choque, el punto de impacto inicial, los daños en el vehículo, y si hubo daño a propiedad pública.

Exploración de los datos

Para iniciar la exploración de datos es necesario comprender qué tipo de datos tiene la muestra por ese motivo se imprimen los cinco primeros datos como se observa en la figuras 1.2,1.2 y 1.3, de esto podemos observar como la gran mayoría de datos en el dataset de arrestos en la ciudad de NY tienen valores cualitativos categóricos o nominales, por lo que están expresados como valores numéricos codificados que hacen referencia a una categoría o grupo en específico.

	summary	ARREST_KEY	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE
1	count	226872	226870	226872	226855	226872	226872
2	mean	2.7064792484001 11E8	424.7544011989245	null	249.3451323532653	null	null
3	stddev	5304010.298148577	274.4753806048603	null	147.68673264760508	null	null
4	min	261180920	1	(null)	101	(null)	(null)
5	max	279779734	997	WEAPONS,MFR,TRANSPORT,ETC.	995	VEHICLE AND TRAFFIC LAWS	VTL21300A5

Imagen 1.1 recorte muestra de datos

LAW_CAT_CD	ARREST_BORO	ARREST_PRECINCT	JURISDICTION_CODE	AGE_GROUP	PERP_SEX	PERP_RACE	X_COORD_CD
225273	226872	226872	226872	226872	226872	226872	226872
9.0	null	63.43052910892486	0.9285367960788462	null	null	null	1005786.72871927
0.0	null	34.635045257003966	7.538568508006557	null	null	null	21509.4376481518
(null)	B	1	0	18-24	F	AMERICAN INDIAN/ALASKAN NATIVE	0
V	S	123	97	<18	U	WHITE HISPANIC	1067220

Imagen 1.2 recorte muestra de datos

Por esta razón ningún dato puede tomarse como cuantitativo, y debido a esta codificación también es más complicado hacer un análisis sin saber cómo decodificar estos datos para

saber el significado de este tipo de datos más allá de la información dada por el diccionario, sin embargo la información de los datos de georeferenciación si pueden ser manipulados dado que esto es una convención internacional que puede ser fácilmente entendida y manipulada.

X_COORD_CD	Y_COORD_CD	Latitude	Longitude	New Georeferenced Column
226872	226872	226872	226872	226872
1005786.7287192778	208289.0843206742	40.7381536574416	-73.92191484770285	null
21509.437648151856	29744.7188726473	0.1182365542455639	0.17333780170454163	null
0	0	0.0	-74.253256	POINT (-73.70059684703173 40.7390218775969)
1067220	271819	40.912714	0.0	POINT (0 0)

Imagen 1.3 recorte muestra de datos

Con los datos de latitud, longitud y georeferenciación se realiza el siguiente mapa de calor, Imagen 2, el objetivo es observar si existen zonas de concentración de arrestos, como es la dispersión de los arrestos por la ciudad para poder identificar si existen patrones que permitan entender la causas de estos problemas.

Como se observa en la Imagen 2 hay una distribución bastante uniforme de forma que si bien hay puntos de calor localizados, los reportes de arrestos están distribuidos por toda la ciudad, sin embargo con los puntos de calor más altos, se va a comenzar a buscar algún tipo de descripción del lugar que permita conocer el por que de estos niveles en esta zonas, una de las razones puede ser la cantidad de población flotante, por que se conoce también que la ciudad de NY para el año 2023 tuvo 61.8 millones de turistas[4]

Imagen 2 mapa de calor de arrestos en la ciudad de NY. Fuente, elaboración propia

Para la información que nos muestra el gráfico de la Imagen 3 se puede denotar la distribución de arrestos en los diferentes boroughs de Nueva York, destacando significativamente a Brooklyn con la mayor cantidad de arrestos, todo esto se puede inferir por la existencia bien sea de una mayor concentración de actividad delictiva, una mayor presencia policial o una combinación de ambas. Por otro lado, para Staten Island se observa una incidencia criminal más baja, que puede darse por menores actividades de vigilancia o una menor incidencia de los delitos, estos datos para los encargados de la formulación de

políticas de seguridad y asignación de recursos podrían justificar una redistribución del personal o una revisión de las estrategias de prevención del crimen. Las autoridades también podrían investigar las causas subyacentes de estas disparidades, como diferencias socioeconómicas, la eficacia de las iniciativas comunitarias de prevención del crimen, la cantidad de población flotante o presencia de turistas.

En el gráfico de la Imagen 3, se puede observar una distribución desigual de arrestos en los diferentes boroughs de Nueva York. Brooklyn destaca significativamente con la mayor cantidad de arrestos, lo que puede atribuirse a diversos factores, como una mayor concentración de actividad delictiva o una mayor presencia policial así como una presencia de ambos factores. Por otro lado, Staten Island presenta una incidencia criminal más baja, que puede deberse a menores actividades de vigilancia o una menor incidencia de delitos.

Para los encargados de la formulación de políticas de seguridad y asignación de recursos, estos datos podrían justificar una redistribución del personal policial o de seguridad, del mismo modo también una revisión de las estrategias de prevención del crimen. Las autoridades también pueden investigar las causas subyacentes de estas disparidades, como diferencias socioeconómicas, la eficacia de las iniciativas comunitarias de prevención del crimen, la cantidad de población flotante o presencia de turistas, que hace una zona objetivo para los delincuentes.

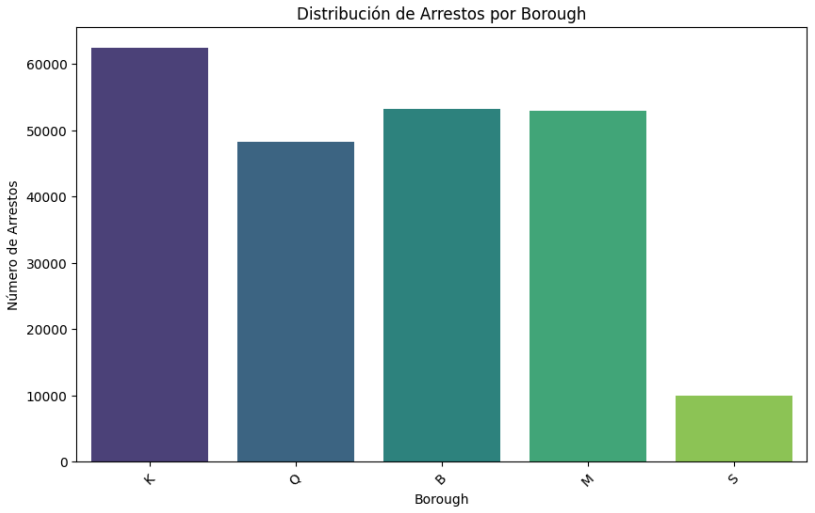


Imagen 3 distribución de arrestos por Borough. Fuente, elaboración propia

Al analizar detenidamente los datos de la columna de fecha graficados en la Imagen 4, se observa un patrón consistente en el comportamiento de los arrestos a lo largo del tiempo. A pesar del transcurso de los meses, se evidencia una estabilidad que sugiere que estos arrestos se mantienen dentro de un cierto "umbral". Esta tendencia indica una falta de evolución significativa en las medidas implementadas durante el año 2023 para abordar esta problemática. Es por esto que es importante analizar a fondo las estrategias actuales y considerar la necesidad de ajustes o nuevas iniciativas con el uso de datos como se hace en este proyecto para lograr un impacto más efectivo en la reducción de los arrestos y mejorar la seguridad pública.

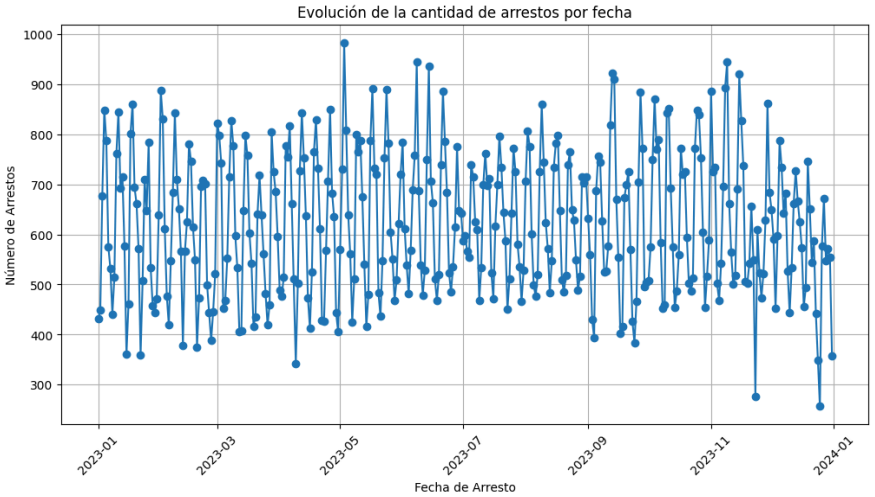


Imagen 4 distribución de arrestos por fecha. Fuente, elaboración propia

Motor Vehicle Collisions - Vehicles

En el ámbito de los siniestros viales, se observaba un incremento constante a lo largo de los años, hasta que la pandemia del COVID-19 irrumpió en el año 2020, momento en el que se produjo una marcada disminución. Este descenso significativo puede atribuirse a las medidas de confinamiento y restricciones implementadas para contener la propagación del virus. En el caso de Estados Unidos, estas medidas se extendieron aproximadamente durante un año y medio, lo que impactó notablemente en la movilidad y la cantidad de vehículos en las carreteras.

Es relevante destacar que esta tendencia a la baja en los siniestros viales se ha mantenido constante desde entonces, con una clara inclinación hacia la reducción. Una de las posibles explicaciones de esta estabilidad positiva podría estar relacionada con la implementación de la nueva ley, Capítulo 229 de 2022, que autoriza la aplicación de multas por exceso de velocidad mediante radares las 24 horas del día. Esta medida representa un cambio significativo respecto al horario anterior, limitado de 6:00 a.m. a 10:00 p.m., lo que ha permitido una mayor vigilancia y control en las vías en todo momento.[5]

Entonces parece ser que la combinación de factores como el impacto inicial de la pandemia, las medidas restrictivas temporales y la implementación continua de normativas más estrictas en materia de seguridad vial ha contribuido a mantener esta tendencia descendente en los accidentes viales.

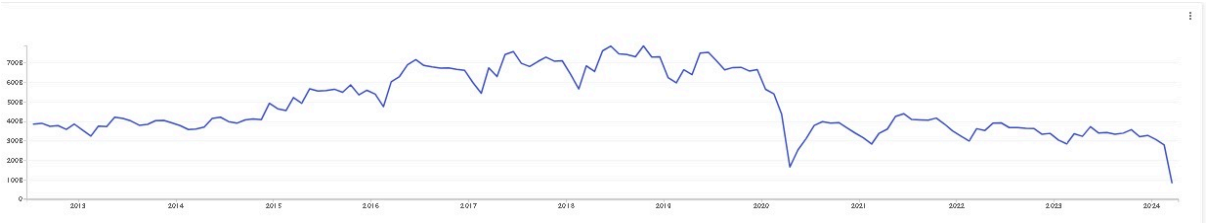


Imagen 5 comportamiento de los accidentes de tránsito en el tiempo Fuente: Elaboración propia

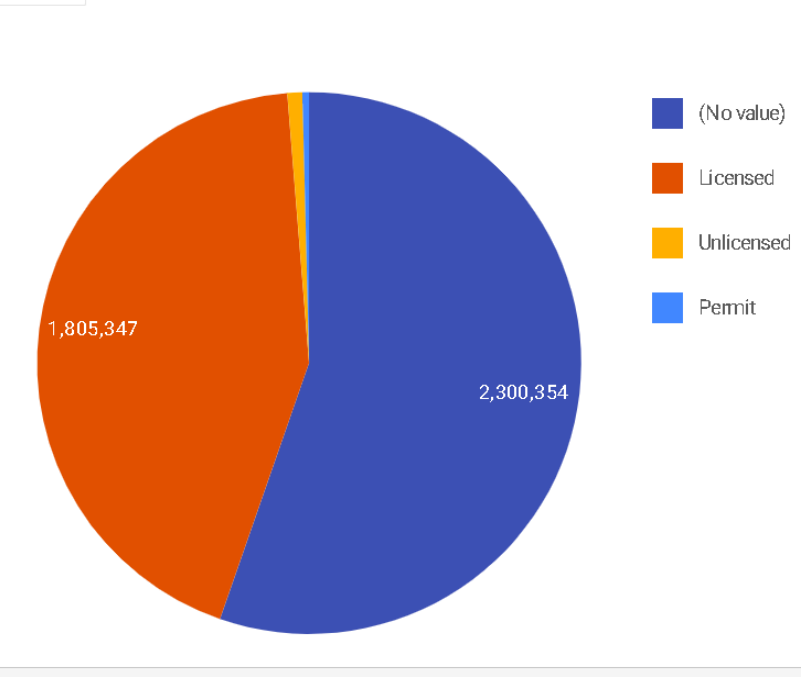


Imagen 6 Estado de la licencia de conducir en los accidentes de tránsito

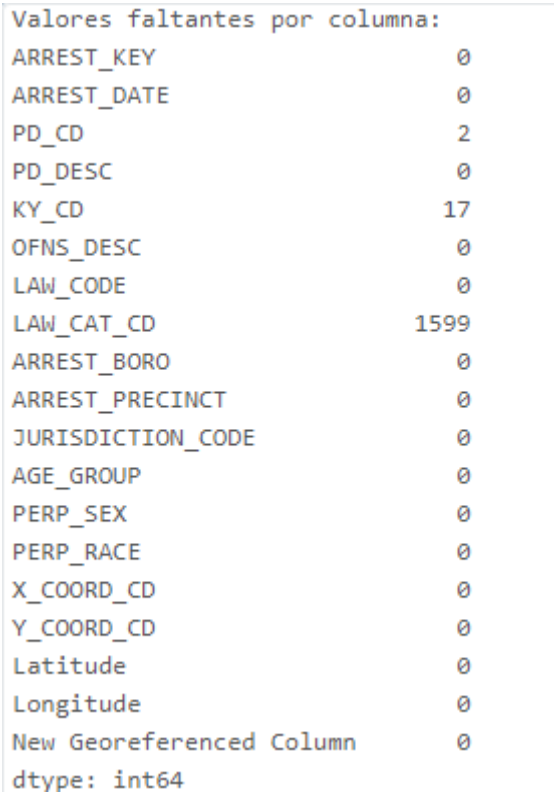
Al observar la imagen 6, llamó la atención un problema en la calidad de los datos, ya que el valor más representativo era *"No value"*. Este hallazgo plantea una interrogante sobre la precisión de los datos recopilados y la eficacia de su captura. No obstante, al analizar la cantidad de personas involucradas en accidentes de tránsito sin licencia, se observa que este número no es particularmente alto.

Este dato es relevante, ya que indica que la mayoría de los accidentes involucran a individuos que sí poseen licencias para conducir. Esto puede sugerir que, aunque la cantidad de personas sin licencia involucradas en accidentes no es significativa para el análisis, por lo cual no presenta mucha relevancia y es necesario revisar otro tipo de valores, por que se esperaba que una población sin licencia que maneja automóviles representan una parte importante de los accidentes de tránsito.

Reporte de calidad de datos

NYPD Arrest Data (Year to Date)

En el presente análisis de la calidad de los datos, se destaca una observación: la columna que ostenta la mayor cantidad de valores faltantes es LAW_CAT_CD, la cual se vincula con el nivel de delito (ya sea felonía, delito menor o infracción). En total, se contabilizan 1599 datos faltantes, lo que representa, en términos cuantitativos, un valor muy pequeño del 0.007% de la muestra, que es casi insignificante. Donde cabe resaltar que, al tratarse de una variable categórica no continua ni discreta, de la cual no se puede hallar ninguna valor estadístico como la media o la moda, para poder se imputada o reemplazada, pero como se mencionó previamente la magnitud de los valores ausentes no adquiere relevancia significativa en el contexto porcentual.



```
Valores faltantes por columna:
ARREST_KEY                    0
ARREST_DATE                   0
PD_CD                         2
PD_DESC                       0
KY_CD                        17
OFNS_DESC                     0
LAW_CODE                      0
LAW_CAT_CD                 1599
ARREST_BORO                   0
ARREST_PRECINCT               0
JURISDICTION_CODE             0
AGE_GROUP                     0
PERP_SEX                      0
PERP_RACE                     0
X_COORD_CD                    0
Y_COORD_CD                    0
Latitude                      0
Longitude                     0
New Georeferenced Column      0
dtype: int64
```

Imagen 7 Calidad de los datos de Arresto

Para abordar la gestión de datos faltantes en el análisis, se presentan diversas estrategias que pueden ser implementadas. Una opción consiste en considerar métodos como la eliminación de filas con valores faltantes, la imputación de la moda (valor más frecuente) o incluso la aplicación de algoritmos de machine learning, como clustering, para agrupar y asignar valores a los datos faltantes en las columnas LAW_CAT_CD, KY_CD y PD_CD. Esta situación se observa de manera similar en las columnas KY_CD y PD_CD.

Sin embargo, al evaluar el enfoque del análisis y la presencia de la columna OFNS_DESC, que ofrece categorías específicas sin valores nulos, surge la posibilidad de eliminar las columnas con datos faltantes (LAW_CAT_CD, KY_CD y PD_CD) para simplificar el conjunto de datos y concentrarse en aquellas variables que proporcionan información completa y relevante para el estudio.

Es fundamental tener en cuenta que la muestra de datos es lo suficientemente amplia como para que, desde una perspectiva estadística, la eliminación de los datos faltantes en las columnas LAW_CAT_CD, KY_CD y PD_CD no tenga un impacto significativo en la representatividad de la muestra. Por lo tanto, eliminar estas columnas con valores faltantes no comprometería la integridad ni la validez de los resultados del análisis. Esta acción permitiría simplificar el conjunto de datos y dirigir el análisis hacia variables completas y pertinentes para el estudio en cuestión.

Además, al considerar el umbral de significancia comúnmente utilizado de 0.05 (5%), donde si la probabilidad de obtener resultados bajo la hipótesis nula es menor al 5%, se considera estadísticamente significativo, se puede concluir que el 0.007% es considerablemente menor que este umbral. Por lo tanto, este porcentaje insignificante no afecta la conclusión general del análisis.

OFNS_DESC
OTHER TRAFFIC INFRACTION
FELONY SEX CRIMES
OTHER OFFENSES RELATED TO THEF
VEHICLE AND TRAFFIC LAWS
KIDNAPPING & RELATED OFFENSES
OFF. AGNST PUB ORD SENSBLTY &
FELONY ASSAULT
ALCOHOLIC BEVERAGE CONTROL LAW

Imagen 8 valores que toma el dato OFNS_DESC

Motor Vehicle Collisions - Vehicles

Para abordar el desafío de calidad de datos en el dataset de collisions es crucial destacar la presencia significativa de valores nulos en diversas categorías. De 21 categorías sólo 5 presentan datos completos, de la misma forma en la mayoría de categorías los datos faltantes son superiores a un millón de registros.

UNIQUE_ID	0
COLLISION_ID	0
CRASH_DATE	0
CRASH_TIME	0
VEHICLE_ID	0
STATE_REGISTRATION	299164
VEHICLE_TYPE	232911
VEHICLE_MAKE	1874569
VEHICLE_MODEL	4098613
VEHICLE_YEAR	1893963
TRAVEL_DIRECTION	1665649
VEHICLE_OCCUPANTS	1778286
DRIVER_SEX	2209212
DRIVER_LICENSE_STATUS	2296864
DRIVER_LICENSE_JURISDICTION	2291944
PRE_CRASH	918633
POINT_OF_IMPACT	1698458
VEHICLE_DAMAGE	1722396
VEHICLE_DAMAGE_1	2587669
VEHICLE_DAMAGE_2	2974293
VEHICLE_DAMAGE_3	3249573
PUBLIC_PROPERTY_DAMAGE	1528858
PUBLIC_PROPERTY_DAMAGE_TYPE	4124336
CONTRIBUTING_FACTOR_1	146119
CONTRIBUTING_FACTOR_2	1685347

dtype: int64

Imagen 9 valores faltantes de Motor Vehicle Collisions

Antes de determinar cómo manejar estos datos faltantes, es fundamental comprender la naturaleza de cada columna, su relevancia para los objetivos del análisis y el tipo de datos que contienen, dado que se puede omitir datos, así como utilizar las técnicas previamentes descritas para imputar datos faltantes en nuestro dataset.

Por ejemplo, la columna PUBLIC_PROPERTY_DAMAGE_TYPE presenta casi la misma cantidad de valores nulos que de filas, sin embargo, al no ser una variable que aporte insights

significativos para analizar el comportamiento de los accidentes o sus causas, se sugiere eliminar esta fila para mantener la integridad del análisis. En contraste, Contributing_Factor_1 es una columna relevante que proporciona información crucial sobre las razones detrás de un accidente.

```
|Accelerator Defective                |
|Driverless/Runaway Vehicle           |
|Unsafe Speed                         |
|Oversized Vehicle                    |
|Passing or Lane Usage Improper       |
|Lane Marking Improper/Inadequate     |
|NULL                                 |
|Aggressive Driving/Road Rage         |
|Other Vehicular                      |
|Driver Inexperience                  |
|Unspecified                          |
|Pavement Defective                   |
|Prescription Medication              |
|View Obstructed/Limited              |
|Lost Consciousness                   |
|Reaction to Other Uninvolved Vehicle|
|Reaction to Uninvolved Vehicle       |
|Fell Asleep                          |
+-------------------------------------+
```

Imagen 10 valores que puede tomar la categoría Contributing_Factor_1

Debido a que aproximadamente el 3.5% de los datos en Contributing_Factor_1 son nulos, y esto sí puede representar un cambio estadístico importante por lo que se plantean opciones como la eliminación de filas con valores faltantes, la imputación de la moda o la exploración de técnicas avanzadas como el clustering. Este enfoque se extiende a la mayoría de las categorías en el dataset, con excepción de la variable VEHICLE_OCCUPANTS. Dado que esta variable es cualitativa, se puede considerar la imputación utilizando medidas como la media, mediana o incluso algoritmos de aprendizaje automático como regresión, KNN u otros modelos predictivos para completar los datos faltantes con mayor precisión.

Sin embargo, vale la pena aclarar que debido a la amplia cantidad de datos atípicos, el realizar una imputación por media, podría generar una imputación de valores que no son del todo correctos, por lo mismo se podría recomendar el mejor imputar por la mediana

```
▸ outliers: pyspark.sql.dataframe.DataFrame = [UNIQUE_ID: integer, COLLISION_ID: integer ... 23 campos adicionales]
Número de valores atípicos en la columna VEHICLE_OCCUPANTS: 875632
```

Por lo tanto, adoptando un enfoque estratégico y detallado para abordar los valores nulos en el dataset, se garantiza una mayor fiabilidad y validez en los resultados del análisis propuesto, permitiendo así extraer conclusiones más sólidas y significativas para generar información, tomar decisiones y generar ideas más valores que aporten a la solución de los problemas que se presentan frente a los accidentes de tránsito.

Planteamiento de preguntas sobre los datos

1. ¿Cuáles serán las zonas de Nueva York con las tasas más altas de arrestos por cada 10,000 habitantes?
2. ¿Qué relación se presenta en cuanto a los arrestos por zona geográfica por conducción o accidente y sus acciones previas o detonantes?

3. ¿Qué áreas con características similares en cuanto a tasas de arresto requieren de una mayor intervención?
4. ¿Cuál es el impacto de variables como la edad, ubicación y raza en la probabilidad de ser arrestado por un tipo específico de crimen?
5. ¿En qué ubicaciones hay una mayor probabilidad de que haya un responsable del arresto específico?.
6. ¿Existe una diferencia notable en el número de arrestos entre los diferentes días de la semana en Nueva York?
7. ¿Cuales son los días entre semana que más choques se presentaron, y los rangos de horas?
8. ¿Qué tipo de vehículos presentan más accidentes, cuantos ocupantes llevan en promedio?
9. ¿Cómo se espera que sea el comportamiento de las colisiones en un futuro cercano?
10. ¿Qué factores contribuyen más a un accidente basado en el clima frecuente por ese tiempo?

Filtros, limpieza y transformación inicial

NYPD Arrest Data (Year to Date)

El proceso de limpieza del dataset "NYPD_Arrest_Data__Year_to_Date__20240312.csv" involucra varias etapas clave para preparar los datos para análisis posteriores:

Carga del Dataset

Inicialmente, se carga el archivo CSV usando Spark, una librería de Python diseñada para el manejo y análisis de datos, recomendada en el proyecto.

Revisión Inicial:

Se examinan las columnas del dataset para entender su estructura, se verifica el tipo de datos de cada columna para identificar posibles inconsistencias, se identifican y cuentan los valores nulos o aquellos equivalentes a "(null)", reemplazándolos por pd.NA para una representación uniforme de los datos faltantes.

Limpieza de columnas: Se eliminan múltiples columnas consideradas no relevantes para el análisis futuro. Estas incluyen:

Códigos y descripciones específicas del arresto (PD_CD, KY_CD, PD_DESC).

La categoría de la ley (LAW_CAT_CD), debido a la falta de uniformidad o relevancia para el análisis propuesto.

Clarificación de Datos:

Los valores en la columna "ARREST_BORO", que originalmente contenían abreviaturas de los nombres de los boroughs de Nueva York, son reemplazados por los nombres completos para mejorar la legibilidad y comprensión de los datos.

De manera similar, los valores en la columna "JURISDICTION_CODE" son reemplazados por términos más descriptivos ("Patrol", "Transit", "Housing", y "Other") para facilitar la interpretación. Esto incluye el mapeo de códigos específicos a la categoría "Other".

Eliminación de Filas con Valores Nulos: Finalmente, se eliminan todas las filas que contienen al menos un valor nulo para asegurar que el dataset resultante esté completo para análisis subsiguientes. Dado que después de hacer limpieza a las columnas solo quedaron 17 registros con datos nulos (226872 en total), se optó por hacer eliminación de estos registros sabiendo que esto no tendría mayor repercusión sobre el estudio.

Motor Vehicle Collisions - Vehicles

Para este primer acercamiento con el tratamiento de este dataset, se realizará la limpieza de dos columnas, en primer lugar se realizará la imputación de los valores nulos a VEHICLE_OCCUPANTS con la mediana de la columna entera, en parte decidido dado a los valores atípicos que se encuentran en gran cantidad en el dataset, pero debido a que como sabemos no pueden haber por ejemplo 3,2 pasajeros, se redondeará está al entero más cercano. Además con la columna CONTRIBUTING_FACTOR_1 al ser la que menos nulos tiene 3,5% del total de los datos, se realizará una imputación por medio de la moda para poder realizar en un futuro una correcta categorización de los valores

Bibliografía

[1]Brookings. (s/f). The Unequal Burden of Crime and Incarceration on America's Poor. Brookings Institution. Recuperado el 2 de febrero de 2024, de https://www.brookings.edu/articles/the-unequal-burden-of-crime-and-incarceration-on-americas-poor/

[2]Oficina del Controlador de Nueva York (OSC), O. del C. del E. (s/f). Economic and Demographic Trends. Oficina del Contralor Del Estado de Nueva York (OSC). Recuperado el 2 de febrero de 2024, de https://www.osc.ny.gov/reports/finance/2023-fcr/economic-and-demographic-trends

[3]Instituto de Investigación en Pobreza (IRP) de la Universidad de Wisconsin-Madison. (s/f). Connections Among Poverty, Incarceration, and Inequality. Instituto de Investigación En Pobreza (IRP) de La Universidad de Wisconsin-Madison. Recuperado el 2 de febrero de 2024, de https://www.irp.wisc.edu/resource/connections-among-poverty-incarceration-and-inequality/

[4]Media, B. (2023 12). tourism-driving-growth-in-new-york-city. Travelpress. https://www.travelpress.com/tourism-driving-growth-in-new-york-city/#:~:text=In%20its%20final%20forecast%20for,the%20city%27s%202018%20visitor%20levels.

[5]The City of New York, DOT, The Office of the New York Mayor. (2023 8). Speeding, Injuries, and Traffic Fatalities Declined in Areas With Speed Cameras During First Year of

24/7 Enforcement. New York City.
https://www.nyc.gov/html/dot/html/pr2023/speed-cameras-first-year.shtml esta referencia