PONTIFICIA UNIVERSIDAD JAVERIANA

Proyecto – Segunda Entrega

Sofia Galindo- Anamaria Leguizamon- Diego Herrera

Procesamiento de Datos a Gran Escala

Ciencia de Datos

Primer Semestre 2024
22/05/2024

Introducción:

Nueva York, una de las ciudades más grandes y vibrantes del mundo, es un crisol de culturas y un epicentro de actividad económica. Con su impresionante arquitectura y un amplia gama de atracciones culturales, desde museos y galerías de arte hasta teatros y eventos musicales, la ciudad atrae a millones de visitantes cada año. Además, su ritmo acelerado y diversidad cultural hacen que siempre haya algo nuevo que ver, hacer o experimentar.

En los últimos años, zonas como Tribeca, Williamsburg, Dumbo o Brooklyn Heights han experimentado una transformación significativa, mejorando la seguridad y dejando atrás su reputación de zonas peligrosas. A pesar de estos avances, la seguridad sigue siendo una preocupación importante. Aunque el 78% de las personas afirman sentirse seguras al caminar solas en la noche, aún se requiere implementar estrategias que disminuyan la incidencia de delitos y accidentes viales.

En 2014, la ciudad de Nueva York implementó el plan Visión Cero con el ambicioso objetivo de eliminar las muertes y lesiones graves por accidentes de tránsito para el año 2024. Este plan ha tenido un impacto positivo, logrando una reducción del 25% en las muertes por accidentes de tránsito entre 2014 y 2023. Entre las estrategias clave del plan se encuentran la reducción de la velocidad vehicular, la ampliación de infraestructura para ciclistas y peatones, la educación vial y la aplicación de las normas de tránsito. Para reducir el indicador de arrestos y accidentes viales, es importante considerar el contexto económico de la ciudad.

El objetivo principal de este proyecto es analizar los datos de arrestos y colisiones vehiculares proporcionados por el Departamento de Policía de Nueva York para obtener insights valiosos que puedan contribuir a la mejora de la seguridad pública y la eficiencia en la administración de justicia. Al comprender las tendencias y patrones en los delitos y accidentes, buscamos responder a preguntas críticas que puedan influir en la toma de decisiones y en la implementación de políticas de seguridad.

Resumen de la Primera Entrega:

En la primera entrega del proyecto, se llevaron a cabo diversas operaciones con el objetivo de garantizar la calidad y consistencia de los datos proporcionados por el Departamento de Policía de Nueva York. El propósito principal fue explorar y preparar los datos para su posterior análisis y obtener una comprensión profunda del problema de negocio. A continuación, se detallan las principales actividades realizadas:

1. Eliminación de Registros Duplicados:
Se procedió a eliminar los registros duplicados en el conjunto de datos. Esta acción garantizó que cada registro en el DataFrame fuera único y no se contabilizará más de una vez en los análisis posteriores.

2. Identificación de Valores Nulos:
Se contabilizaron los valores nulos presentes en cada columna de los Data Frame utilizando las funciones `count` y `when` de PySpark. Este paso permitió identificar las columnas que contenían valores faltantes y tomar las medidas apropiadas para su tratamiento.

3. Imputación de Valores Nulos:
Para abordar los valores nulos, se aplicó una estrategia de imputación en las columnas

`PD_CD` y `KY_CD`, reemplazando los valores faltantes por la etiqueta 'Desconocido'. Esta técnica evitó la pérdida de registros completos y permitió mantener la integridad de los datos.

Entrega 2: Preparación de datos, modelado y presentación de resultados.

Filtros y transformaciones:

En esta etapa primera etapa de la segunda entrega, se realizaron diversas transformaciones finales y se aplicaron filtros a los datos previamente tratados, con el fin de prepararlos para el análisis y garantizar su máxima calidad y consistencia. En primer lugar, se identificaron y eliminaron columnas innecesarias de los dataframes de arrestos y accidentes viales, esto basado en las preguntas planteadas en la primera entrega.

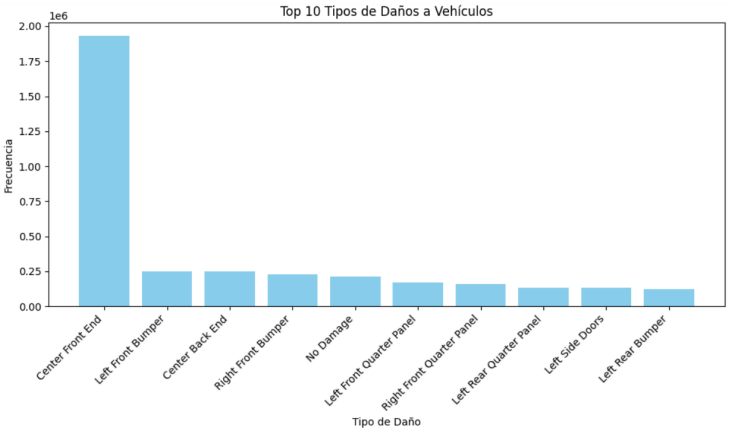
Además, se abordó el manejo de los valores nulos en la columna LAW_CAT_CD en el dataframe de arrestos y las columnas VEHICLE_TYPE y PRE_CRASH en el dataframe de accidentes vehiculares. Para el dataframe de arrestos, se implementó una estrategia de imputación basada en la moda para la columna LAW_CAT_CD, utilizando AGE_GROUP y PERP_SEX como variables de partición. En cuanto al dataframe de accidentes viales, se aplicó un enfoque similar para imputar los valores faltantes en las columnas VEHICLE_TYPE y PRE_CRASH. En el primer caso, se calculó la moda de VEHICLE_TYPE por VEHICLE_MODEL, mientras que para PRE_CRASH se utilizó VEHICLE_DAMAGE como variable de partición. Estos procesos permitieron asignar los valores más frecuentes a los registros con datos faltantes, evitando la pérdida de información valiosa.

Por otra parte, se realizaron transformaciones en las columnas de fechas para facilitar el análisis temporal. Las columnas `ARREST_DATE` y `CRASH_DATE` fueron convertidas al formato de fecha adecuado, y se extrajeron los años y meses correspondientes en nuevas columnas, permitiendo un examen más detallado de los patrones temporales.

Finalmente, se optimizó el manejo de la información geográfica al combinar las coordenadas de latitud y longitud en un solo campo denominado `LOCATION`. Esta integración de datos facilitará el análisis espacial y la visualización de patrones geográficos relacionados con los arrestos y accidentes viales.

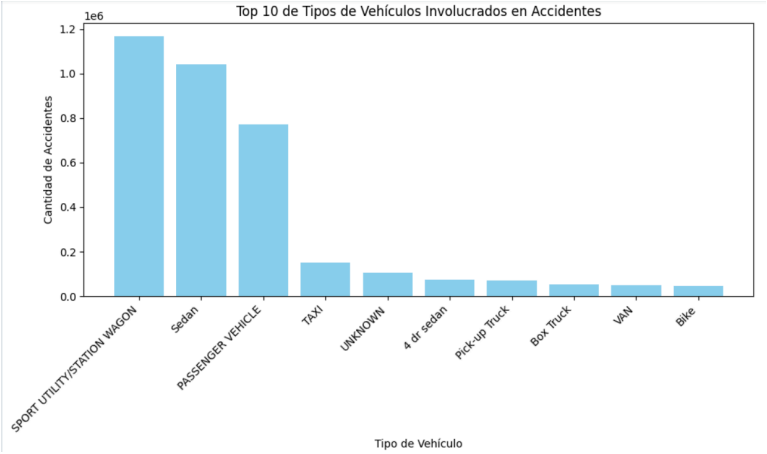
Respuesta a preguntas de negocio planteadas:

a. ¿Es frecuente que los autos tengan daños en lugares específicos (VEHICLE_DAMAGE) después de un accidente?



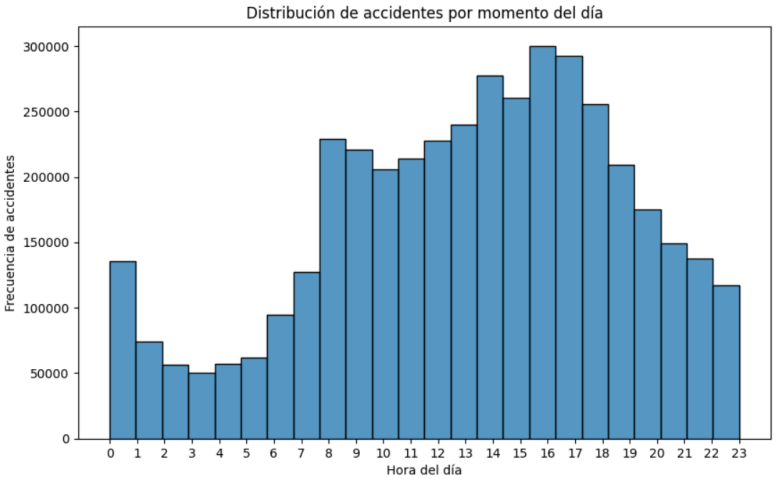
El gráfico evidencia que los daños en la parte frontal central de los vehículos son los más frecuentes después de un accidente. La categoría "Center Front End" exhibe una frecuencia que supera a todas las demás, dejando atrás incluso a las siguientes categorías más comunes como daños en los bumpers y cuartos delanteros. Esta diferencia sugiere que los impactos frontales directos representan el escenario de colisión por excelencia en los siniestros viales reportados. Esta tendencia predominante en los patrones de daño podría atribuirse a factores como el tráfico bidireccional, accidentes en intersecciones, colisiones con objetos fijos, entre otros escenarios donde es más factible un choque frontal.

b. ¿Existe un patrón de accidentes relacionado con modelos específicos de vehículos (VEHICLE_MODEL)?



La distribución de accidentes por tipos de vehículos involucrados exhibe un claro patrón dominado por los SUV y sedanes. Estos dos modelos sobresalen de forma abrumadora con una frecuencia que supera con creces al resto de categorías vehiculares. Los automóviles clasificados genéricamente como "PASSENGER VEHICLE" también registran niveles considerables, aunque por debajo de los SUV y sedanes. En contraste, otras modalidades como taxis, camiones, furgonetas y bicicletas tienen una representación notablemente menor en los siniestros reportados. Esta tendencia podría atribuirse en gran medida a la alta prevalencia de SUV y sedanes en las vías, dados su popularidad como vehículos de uso personal y familiar. Su extendida presencia en el tráfico diario incrementaría naturalmente la probabilidad de verse involucrados en más accidentes en comparación con modalidades menos numerosas.

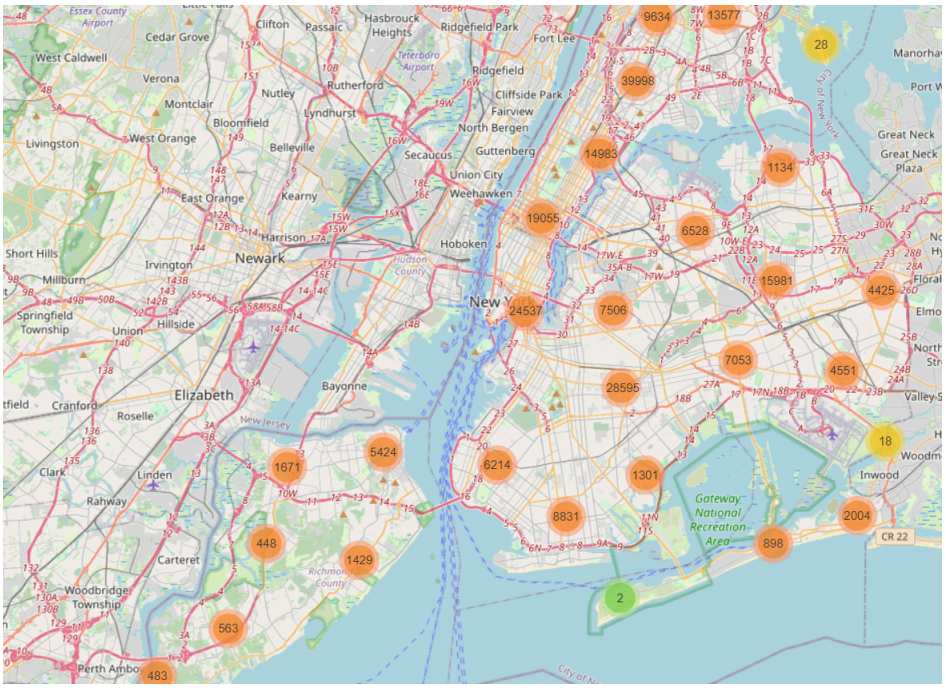
c. ¿Los vehículos involucrados en accidentes (VEHICLE ID) presentan patrones específicos en las fechas y horas de los choques (CRASH DATE, CRASH TIME) que podrían ayudar a prevenir futuros incidentes?



La visualización de la distribución horaria de accidentes revela patrones interesantes. Se distinguen dos picos en la frecuencia de siniestros: uno en las primeras horas de la mañana, entre las 7 y las 8 am, y otro por la tarde, entre las 4 y las 5 pm. Esta tendencia bimodal podría estar fuertemente vinculada a los períodos de máxima congestión vehicular en la ciudad, cuando millones de conductores se desplazan hacia y desde sus lugares de trabajo o estudio. El aumento drástico del tráfico durante estas "horas pico" incrementa exponencialmente las interacciones vehículo-vehículo y vehículo-peatón, elevando por consiguiente el riesgo de colisiones. Identificar estos períodos críticos es crucial para implementar estrategias efectivas que mitiguen la ocurrencia de accidentes, como reforzar los controles de tránsito, optimizar la señalización y sincronizar adecuadamente los semáforos durante esos lapsos.

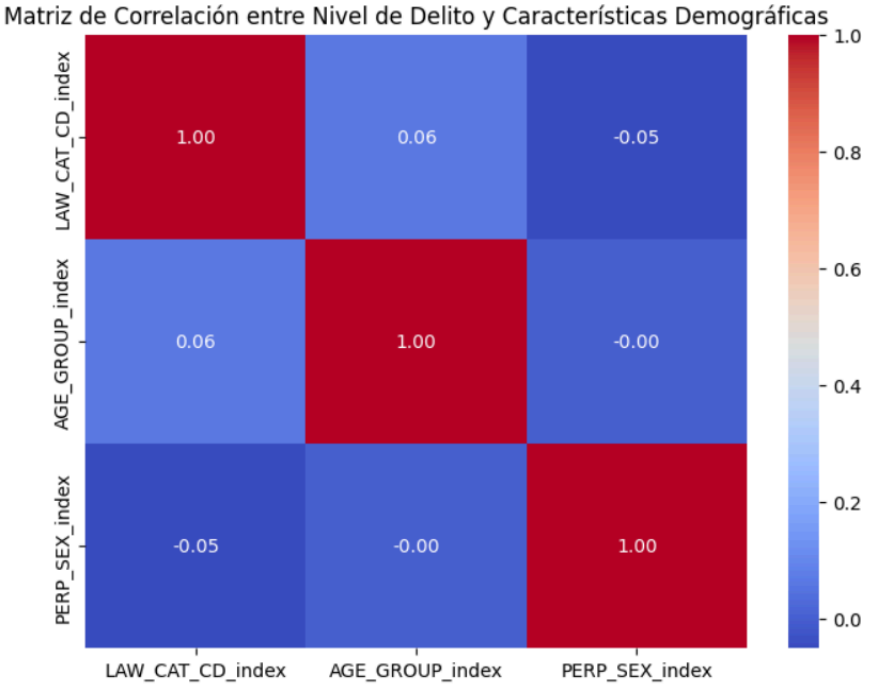
d. ¿Existe un patrón en las acciones que realizaban los vehículos (PRE_CRASH) justo antes de los accidentes?

e. ¿Qué tipos de delitos (*PD DESC* y *OFNS DESC*) son los más frecuentes y cómo se distribuyen geográficamente (*ARREST BORO*, *ARREST_PRECINCT, X_COORD_CD, Y_COORD_CD, Latitud, Longitud*)?



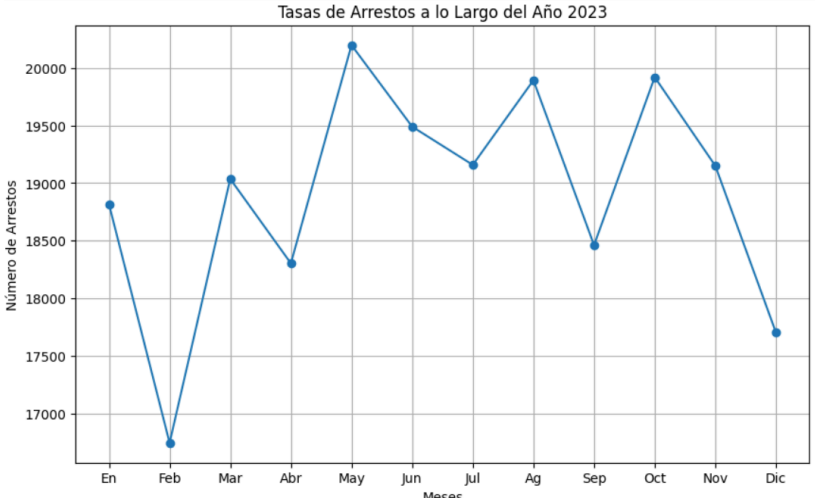
La distribución geográfica de los delitos muestra patrones contundentes. Las zonas con mayor concentración criminal coinciden con áreas urbanas densamente pobladas y con intensa actividad económica y social, como lo evidencian los grandes círculos en Manhattan, Brooklyn y el Bronx. Esta correlación sugiere que la afluencia masiva de personas en lugares como estaciones de transporte, centros comerciales y otros puntos neurálgicos de la ciudad, incrementa las oportunidades delictivas. Por tanto, las autoridades deberían reforzar la vigilancia policial y las medidas preventivas precisamente en estos puntos críticos. Algunas estrategias efectivas podrían incluir aumentar el pie de fuerza policial, instalar más iluminación disuasoria en calles y parques, así como implementar programas comunitarios que fomenten la convivencia pacífica y brinden alternativas prosociales, especialmente entre los jóvenes en riesgo. Optimizar la distribución de los recursos en función de estos patrones delictivos permitiría a los cuerpos de seguridad utilizar sus capacidades de manera más focalizada y eficiente.

f. ¿Existe alguna relación entre el nivel de delito (*LAW CAT CD*) y las características demográficas de los sospechosos, como su grupo de edad (*AGE GROUP*) o género (PERP_SEX))?

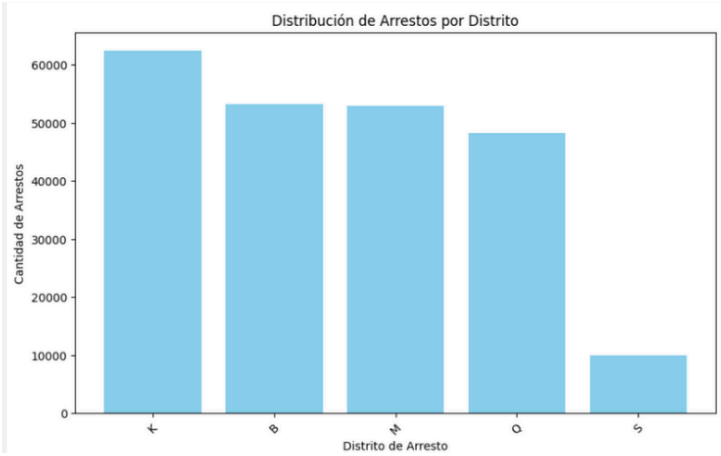


Las correlaciones entre las variables demográficas y el nivel de delito son sumamente débiles según la matriz presentada. La correlación de apenas 0.06 entre LAW_CAT_CD_index y AGE_GROUP_index indica que no existe una relación lineal apreciable entre el tipo de crimen cometido y la edad del perpetrador. De forma similar, la correlación negativa pero diminuta de -0.05 entre LAW_CAT_CD_index y PERP_SEX_index sugiere que el género del criminal tampoco está linealmente asociado al nivel de delito. Incluso la correlación de -0.00 entre AGE_GROUP_index y PERP_SEX_index manifiesta la ausencia total de relación lineal entre la edad y el sexo de los sospechosos. En conjunto, estos valores de correlación próximos a cero no respaldan la existencia de vínculos lineales significativos entre las características demográficas examinadas y la gravedad de los crímenes cometidos dentro de los datos analizados.

g. ¿Cómo han variado las tasas de arrestos a lo largo del tiempo (*ARREST DATE*) y si existen patrones temporales o estacionales en los diferentes tipos de delitos?



h. ¿Qué distritos policiales (*ARREST PRECINCT*) o localidades (boroughs) (*ARREST BORO*) tienen las mayores tasas de arrestos y cuáles son los tipos de delitos predominantes en esas áreas?



ARREST_BORO	LAW_CAT_CD	count	rank
B	M	30878	1
B	F	21805	2
B	V	187	3
K	M	32149	1
K	F	29189	2
K	V	754	3
M	M	29711	1
M	F	21994	2
M	V	245	3
Q	M	26014	1
Q	F	21369	2
Q	V	118	3
S	M	5729	1
S	F	4264	2
S	V	8	3

El examen de las estadísticas de arrestos revela diferencias notables en las tasas y patrones delictivos entre los distintos distritos de Nueva York. Brooklyn destaca como el borough con el mayor volumen total de arrestos, encabezando en todas las categorías principales: delitos menores, delitos graves y violaciones. Este elevado índice de detenciones en Brooklyn puede atribuirse en parte a su alta densidad poblacional y su diversidad socioeconómica, factores que suelen estar correlacionados con mayores niveles de actividad criminal. Sin embargo, la considerable prevalencia de arrestos por violación en este distrito sugiere la necesidad de un enfoque específico para abordar los delitos sexuales y brindar apoyo integral a las víctimas. El contraste con otras áreas como el Bronx y Queens, que exhiben menores niveles de arrestos por violación, podría indicar variaciones significativas en la dinámica subyacente de estos delitos o en el manejo policial y judicial de los mismos.

Selección de técnicas de aprendizaje de máquina.

Aprendizaje Supervisado: Regresión Logística Multinomial

- Se fundamenta en la naturaleza de los datos y el tipo de relación que se busca modelar entre las variables predictoras y la variable objetivo. En el contexto de nuestro análisis sobre los arrestos policiales y la raza de los individuos implicados, optamos por este método debido a la diversidad y la independencia entre las distintas categorías raciales.
- La Regresión Logística Multinomial es la elección adecuada cuando las categorías de la variable dependiente no presentan un orden natural o jerarquía clara entre sí. En nuestro caso, las distintas clasificaciones raciales, como caucásico, afroamericano, latino, asiático, entre otras, no están inherentemente ordenadas en una secuencia lógica.
- Al emplear la Regresión Logística Multinomial, podemos modelar la probabilidad de que un individuo pertenezca a una categoría racial específica en función de múltiples variables predictoras, como el género, la edad, la ubicación, entre otras. Este enfoque nos permite comprender cómo estas variables influyen en la pertenencia a una categoría racial en particular, sin asumir un orden predefinido entre las categorías raciales.

Aprendizaje No Supervisado: K-Means Clustering.

- **Agrupación Natural de Datos:** El K-Means Clustering nos permite identificar patrones y agrupaciones naturales en los datos sin necesidad de etiquetas predefinidas. Esto es especialmente útil en datos de arrestos, donde puede no estar claro de antemano cuáles son los grupos relevantes o significativos.
- **Detección de Perfiles de Arresto:** Utilizando K-Means Clustering, podemos detectar distintos perfiles de arrestos basados en características como edad, género, tipo de delito y ubicación geográfica. Esto puede revelar insights sobre diferentes comportamientos y características de los arrestos que no son evidentes mediante un análisis simple.
- **Simplicidad y Eficiencia:** El K-Means es relativamente simple y computacionalmente eficiente, lo que lo hace adecuado para grandes conjuntos de datos como los registros de arrestos policiales. Su simplicidad también facilita la interpretación de los resultados y la implementación de mejoras iterativas.
- Identificación de Anomalías: Al formar grupos basados en similitudes, el K-Means puede ayudar a identificar anomalías o casos atípicos en los datos de arrestos. Estos outliers pueden ser indicativos de eventos inusuales o comportamientos específicos que merecen una mayor atención.
- **Mejora de la Toma de Decisiones:** La agrupación de datos mediante K-Means puede proporcionar a los responsables de la toma de decisiones una visión más clara de las tendencias y patrones en los arrestos. Esto puede influir en la formulación de políticas, la asi**gnación de recursos y el desarrollo de estrategias de intervención específicas.**
- **Exploración y Segmentación de Datos:** K-Means permite una exploración profunda y una segmentación efectiva de los datos, facilitando el análisis de subgrupos específicos. Por ejemplo, podemos segmentar los arrestos por distrito y tipo de delito para entender mejor las dinámicas locales y adaptar las estrategias de prevención del crimen.

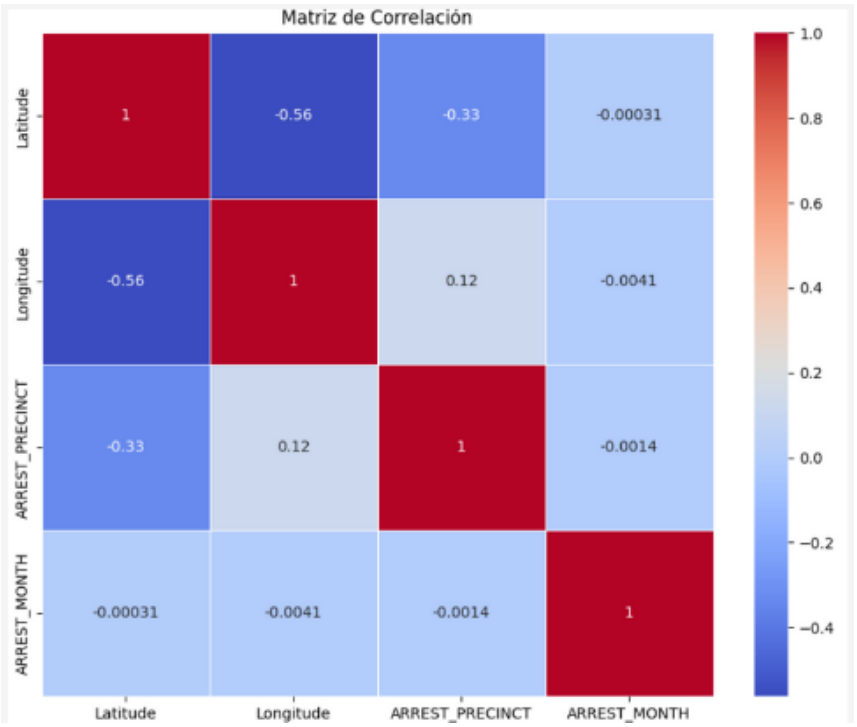
Preparación de datos para modelado.

Datos de arrestos del Departamento de Policía de Nueva York

1. ***Eliminar características fuertemente correlacionada***

No se encontraron características fuertemente correlacionadas gracias al excelente proceso de limpieza de datos realizado anteriormente. Este éxito se debe también al profundo conocimiento del negocio y a la capacidad de mantener los objetivos claros desde las primeras etapas del proyecto. La meticulosa preparación y la atención al detalle en la fase de limpieza de datos permitieron identificar y eliminar cualquier posible sesgo o error. Además, el entendimiento integral de las necesidades y objetivos del negocio garantizó que el enfoque se mantuviera alineado con las metas estratégicas, lo cual facilitó la obtención de resultados precisos y relevantes.

Resultado.



2. Normalización utilizando StandardScaler

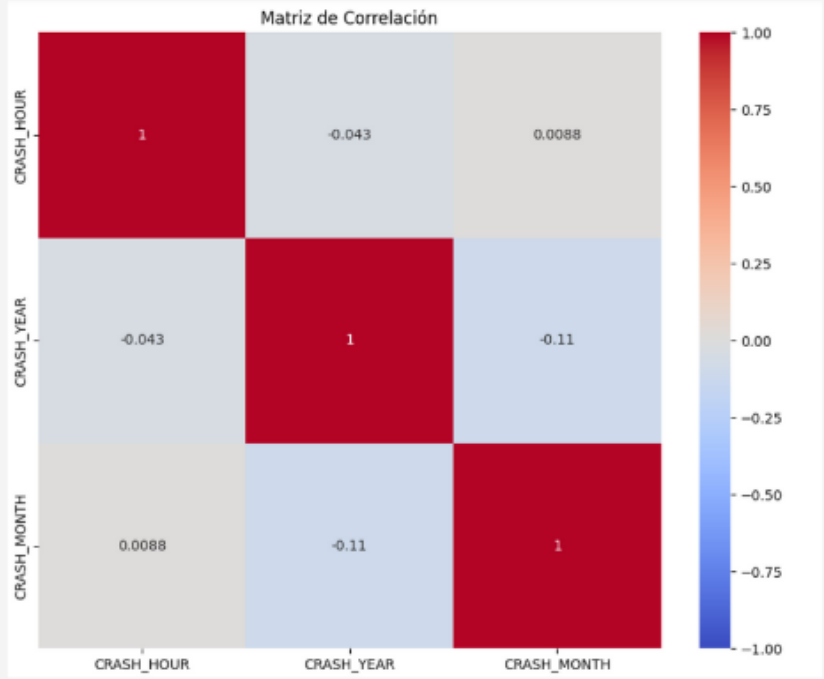
Se llevó a cabo la normalización utilizando el método StandardScaler para evitar que alguna variable numérica tenga mayor peso al momento de construir el modelo. Este proceso asegura que todas las variables numéricas contribuyan de manera equitativa al análisis. No se realizó la normalización de variables categóricas debido al uso de la Regresión Logística Multinomial, la cual maneja de manera adecuada las variables categóricas sin necesidad de normalizarlas. Este enfoque garantiza que el modelo sea robusto y preciso, aprovechando las ventajas de la Regresión Logística Multinomial para el tratamiento de variables categóricas.

Resultado.

Latitude	Longitude	ARREST_PRECINCT	ARREST_MONTH	ARREST_YEAR
0.994949009719784	0.45002790467593384	-0.4166453082952651	-1.6226521044544153	0.0
-0.7878921881233384	-0.8962162365631693	1.6333014861482094	-1.6226521044544153	0.0
-1.2187403325491275	-0.2247123934579678	-0.07017542754425529	-1.6226521044544153	0.0
0.26555529091358127	0.037809685122191586	1.4600665457727044	-1.6226521044544153	0.0
0.7271976345042067	0.027373415697949703	-0.5610077586081857	-1.6226521044544153	0.0
-0.4606245317973604	-0.453872676644313	0.3629119233945069	-1.6226521044544153	0.0
0.8972381108043329	0.47333499614862756	-0.4166453082952651	-1.6226521044544153	0.0
-0.20358896277294034	-0.19469009048182664	0.7671267842706849	-1.6226521044544153	0.0
-0.9273414481699327	-0.11231913700485954	-0.012430447419087002	-1.6226521044544153	0.0
1.0435380441065014	-0.04235170993043134	-0.8497326592340273	-1.329401163492822	0.0
0.8880615916904315	0.08718725836841508	-0.5610077586081857	-1.329401163492822	0.0
-0.10038906763171214	1.0797116720174094	1.200214135209447	-1.329401163492822	0.0
-1.1708363654879794	0.9283367281680055	1.0847241749591106	-1.6226521044544153	0.0
-0.36440217423129406	-0.30692181263759744	0.7093818041455167	-1.6226521044544153	0.0
-0.9273414481699327	-0.11231913700485954	-0.012430447419087002	-1.329401163492822	0.0

Colisiones de vehículos motorizados - Vehículos

1. ***Eliminar características fuertemente correlacionada***



No se encontraron características fuertemente correlacionadas gracias al excelente proceso de limpieza de datos realizado anteriormente. Este éxito se debe también al profundo conocimiento del negocio y a la capacidad de mantener los objetivos claros desde las primeras etapas del proyecto. La meticulosa preparación y la atención al detalle en la fase de limpieza de datos permitieron identificar y eliminar cualquier posible sesgo o error. Además, el entendimiento integral de las necesidades y objetivos del negocio garantizó que el enfoque se mantuviera alineado con las metas estratégicas, lo cual facilitó la obtención de resultados precisos y relevantes.

2. **Normalización utilizando StandardScaler**

Se llevó a cabo la normalización utilizando el método StandardScaler para evitar que alguna variable numérica tenga mayor peso al momento de construir el modelo. Este proceso asegura que todas las variables numéricas contribuyan de manera equitativa al análisis.

Resultado

CRASH_HOUR	CRASH_YEAR	CRASH_MONTH
-0.7227146923115745	-1.644366187577866	0.6796513302524124
-0.8959166386712513	0.6278604120724706	0.6796513302524124
0.6629008785658396	-0.6705547877277217	0.9723472668926723
1.1825067176448698	-0.6705547877277217	0.9723472668926723
1.3557086640045466	-1.319762387627818	-0.7838283529488864
0.6629008785658396	-0.3459509877776737	-0.49113241630862664
-0.02990690687286746	0.6278604120724706	0.9723472668926723
-0.2031088532325442	-0.3459509877776737	0.3869553936121527
-0.37631079959222097	-1.319762387627818	0.0942594569718929
0.8361028249255164	-1.644366187577866	1.265043203532932
0.31649698584648606	-0.3459509877776737	0.0942594569718929
-0.8959166386712513	-0.6705547877277217	0.3869553936121527
1.3557086640045466	-0.3459509877776737	0.3869553936121527
0.6629008785658396	0.6278604120724706	0.9723472668926723
0.8361028249255164	0.9524642120225186	-0.1984364796683669
1.009304771285193	-1.644366187577866	0.3869553936121527
0.1432950394868093	-0.3459509877776737	0.6796513302524124
0.31649698584648606	-1.319762387627818	-0.7838283529488864

Selección de variables según criterio de negocio.

Variable escogida raza del detenido (PERP_RACE)

La selección de la variable "raza del detenido" como variable objetivo se fundamenta en la importancia de comprender y abordar posibles disparidades raciales en el sistema de justicia penal. Esta variable nos permite investigar y analizar la posible influencia de factores raciales en el proceso de arresto, lo que puede arrojar luz sobre posibles sesgos o inequidades en la aplicación de la ley.

Explorar la relación entre la raza de los individuos detenidos y otras variables predictoras nos permite identificar patrones y tendencias que podrían indicar desigualdades sistemáticas en la aplicación de la ley. Además, al comprender cómo la raza puede estar relacionada con el riesgo de arresto o el tipo de cargos enfrentados, podemos avanzar hacia estrategias más equitativas y justas en el sistema de justicia penal.

Al centrarnos en la variable de raza del detenido, buscamos no solo cuantificar la prevalencia de arrestos entre diferentes grupos raciales, sino también examinar las posibles causas subyacentes de las disparidades observadas. Este enfoque nos brinda la oportunidad de promover una mayor transparencia y rendición de cuentas en el sistema de justicia penal y de trabajar hacia una aplicación más justa y equitativa de la ley para todos los ciudadanos.

Técnicas de Machine Learning con ML Lib

Preprocesamiento de Datos
El preprocesamiento es una etapa crucial en cualquier flujo de trabajo de análisis de datos, especialmente en el aprendizaje automático. En este proyecto, comenzamos con la indexación

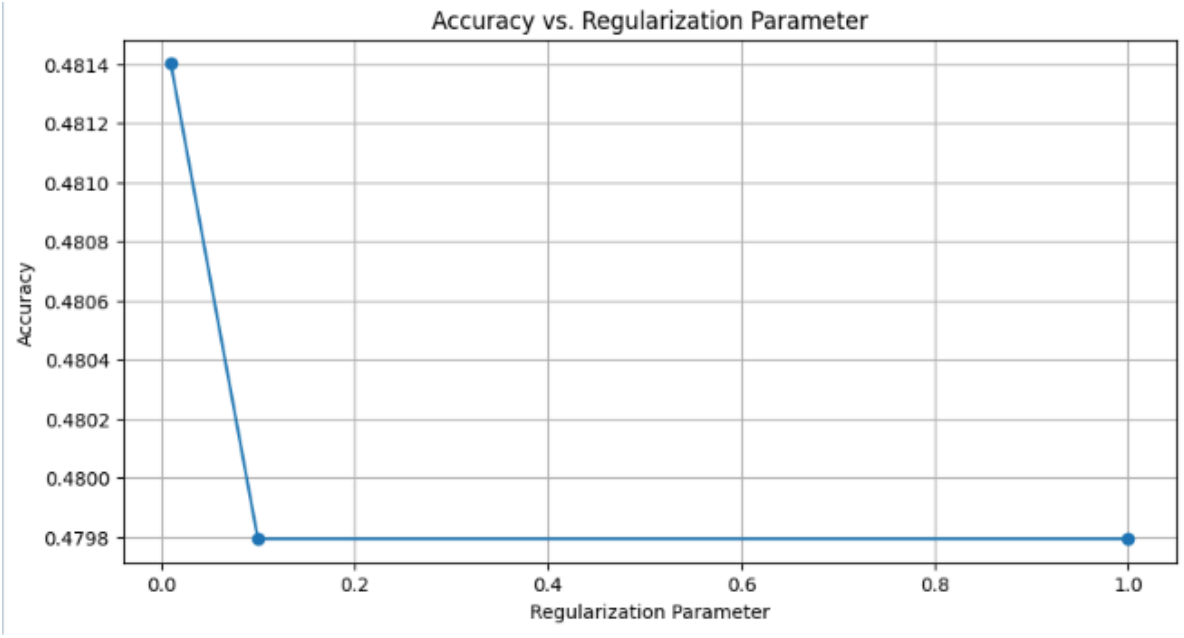
de variables categóricas, específicamente la variable PERP_RACE. Esta conversión es necesaria porque los algoritmos de machine learning generalmente requieren que los datos de entrada sean numéricos. Utilizamos StringIndexer de Spark MLlib para transformar esta variable categórica en índices numéricos.

Posteriormente, procedimos a ensamblar varias características numéricas en un único vector. Esto se logró mediante VectorAssembler, que combina columnas de características como ARREST_PRECINCT, Latitude, Longitude, ARREST_YEAR, y ARREST_MONTH en una única representación numérica. Este vector de características es fundamental, ya que proporciona una representación compacta y eficiente de los datos para el entrenamiento de modelos.

Aprendizaje Supervisado: Regresión Logística

El conjunto de datos procesado se divide en un 70% para entrenamiento y un 30% para pruebas, asegurando una evaluación robusta del modelo. Se configura un modelo de regresión logística, variando el parámetro de regularización en tres niveles diferentes (0.01, 0.1, 1.0) para estudiar su impacto en la precisión del modelo. Cada configuración es evaluada usando un evaluador de clasificación multiclase, que proporciona métricas como la precisión y el recall del modelo.

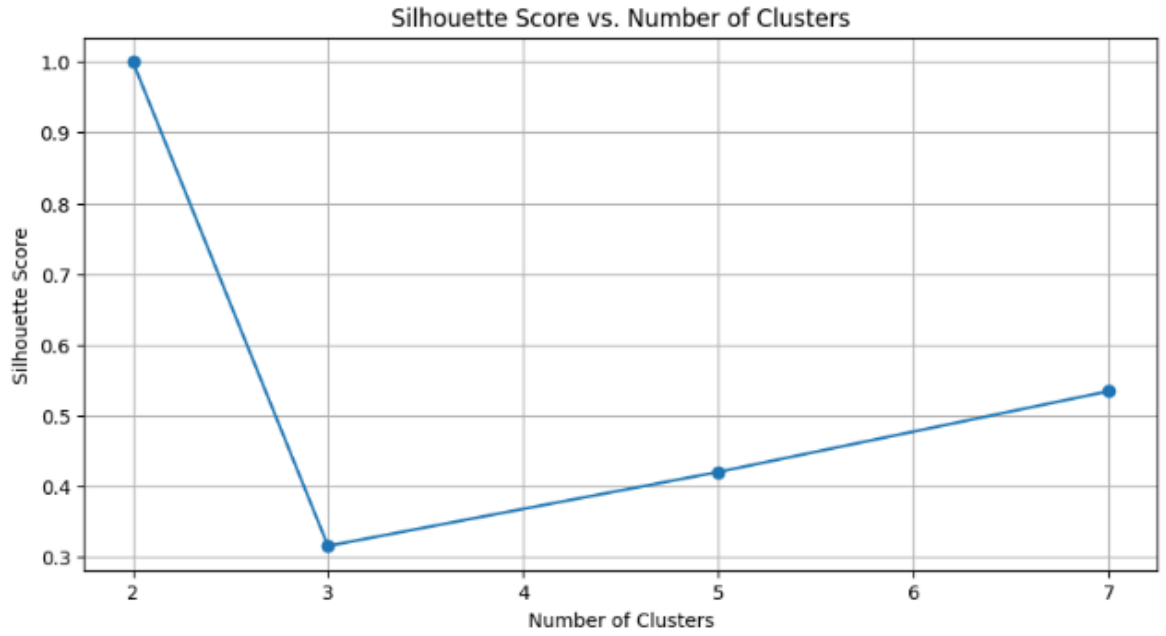
- **Indexación de Variables:** Uso de StringIndexer para convertir la variable categórica PERP_RACE a formato numérico en df1_scaled.
- **Ensamblaje de Características:** Creación de vectores de características con VectorAssembler usando columnas como 'ARREST_PRECINCT', 'Latitude', 'Longitude', 'ARREST_YEAR', 'ARREST_MONTH'.



La gráfica muestra la relación entre diferentes valores de parámetro de regularización (0.01, 0.1, 1.0) y la precisión del modelo de regresión logística. Cada punto en la gráfica representa la precisión obtenida con un valor específico de regularización, mostrando cómo varía la precisión con cambios en el parámetro.

Análisis de Clustering con K-Means

Se utiliza el algoritmo K-means para identificar grupos o clústeres dentro del conjunto de datos, basándose en las características numéricas ensambladas previamente. El número de clústeres varía y se evalúa la calidad de cada configuración utilizando el coeficiente Silhouette, que mide qué tan similares son los datos dentro de un clúster en comparación con otros clústeres. Los resultados de este análisis sugieren diferentes estrategias para la segmentación de clientes o la focalización de recursos en un contexto empresarial.



Variación del Silhouette Score

- Con dos clústeres, el Silhouette Score es casi 1.0, lo que indica una muy buena separación y cohesión dentro de los clústeres. Esto sugiere que los dos clústeres están muy bien definidos y separados entre sí.
- A medida que aumenta el número de clústeres a tres, el Silhouette Score cae drásticamente cerca de 0.3, lo que indica una pobre separación y cohesión. Esto podría interpretarse como que agregar un tercer clúster no contribuye positivamente a la estructura de agrupamiento, posiblemente debido a la superposición significativa entre clústeres o a la distribución irregular de los datos.
- Desde cuatro clústeres en adelante, el Silhouette Score muestra una tendencia ascendente gradual hasta alcanzar aproximadamente 0.5 con siete clústeres. Este incremento sugiere una mejora progresiva en la definición de los clústeres con la adición de más grupos, aunque los valores siguen siendo moderados, lo que podría indicar una estructura de clústering menos clara que con dos clústeres.

Interpretación Estratégica:

- El pico inicial con dos clústeres puede ser ideal si el objetivo es tener una división clara y amplia, útil en situaciones donde se requieren distinciones generales entre grupos grandes de datos.

- El aumento gradual del Silhouette Score con más de cuatro clústeres puede ser útil para detalles más granulares en el análisis, permitiendo identificar subgrupos dentro de los datos que pueden ser de interés para análisis específicos o para dirigir estrategias diferenciadas en áreas como marketing o desarrollo de productos.

Conclusiones

Según los resultados de los modelos y las preguntas planteadas, podemos suponer que:

1. **Posible Sesgo Racial en la Policía de Nueva York**: Los datos sugieren la posibilidad de un sesgo racial por parte de la policía de Nueva York en la realización de arrestos.
2. **Delitos y Factores Socioeconómicos**: Las personas con menor acceso al empleo y poder adquisitivo tienden a cometer más delitos, lo que indica una posible correlación entre la situación económica y la actividad delictiva.
3. **Diferencias Distritales en Problemáticas de Seguridad**: Cada distrito de Nueva York presenta características y problemáticas de seguridad únicas, lo que requiere la implementación de técnicas y políticas específicas para cada uno.
4. **Impacto de las Políticas en Accidentes**: Las políticas implementadas han tenido un gran efecto en la reducción de accidentes en Nueva York, demostrando la efectividad de las medidas de seguridad vial.
5. **Conocimiento y Cumplimiento de las Leyes de Tránsito**: Aunque los conductores muestran un buen conocimiento de las leyes de tránsito, parece haber una intencionalidad en violar estas leyes, posiblemente debido a la percepción de baja probabilidad de ser sancionados.
6. **Gravedad y Velocidad de los Accidentes**: La mayoría de los accidentes no son graves y ocurren a bajas velocidades, lo que sugiere que muchos de estos incidentes podrían ser evitables con una conducción más atenta y prudente.
7. **Relación entre Atascos y Accidentes**: Podría existir una relación entre los atascos de tráfico y los accidentes, posiblemente debido a la impaciencia de los conductores que puede llevar a comportamientos arriesgados.