

Proyecto Entrega 2

Por: Sofía Catalina Galindo, Diego Alejandro Herrera, Anamaría Leguizamón.



Programa

- Filtros y transformaciones
- Respuesta a preguntas de negocio planteada
- Selección de técnicas de aprendizaje de máquina
- Preparación de datos para modelado
- técnicas seleccionadas
- Métricas

Filtros y transformaciones



[Volver a la página de agenda](#)

ELIMINACIÓN DE DATOS

```
▶ ✓ 05:53 PM (1s)

columns_to_drop_vehicle = [
    "UNIQUE_ID",
    "COLLISION_ID",
    "STATE_REGISTRATION",
    "VEHICLE_MAKE",
    "VEHICLE_YEAR",
    "DRIVER_SEX",
    "DRIVER_LICENSE_STATUS",
    "DRIVER_LICENSE_JURISDICTION",
    "POINT_OF_IMPACT",
    "VEHICLE_OCCUPANTS",
    "PUBLIC_PROPERTY_DAMAGE",
    "PUBLIC_PROPERTY_DAMAGE_TYPE",
    "CONTRIBUTING_FACTOR_1",
    "CONTRIBUTING_FACTOR_2",
    "VEHICLE_DAMAGE_1",
    "VEHICLE_DAMAGE_2",
    "VEHICLE_DAMAGE_3",
    "TRAVEL_DIRECTION"
]
```

```
columns_to_drop_arrest = [
    "ARREST_KEY",
    "PD_CD",
    "KY_CD",
    "JURISDICTION_CODE",
    "X_COORD_CD", # Si se usa Latitude y Longitude
    "Y_COORD_CD", # Si se usa Latitude y Longitude
    "New Georeferenced Column"
]
```

DATOS FALTANTES ARRESTOS

|ARREST_DATE|PD_DESC|OFNS_DESC|LAW_CODE|LAW_CAT_CD|ARREST_BORO|ARREST_PRECINCT|AGE_GROUP|PERP_SEX|PERP_RACE|Latitude|Longitude|ARREST_YEAR|ARR
EST_MONTH|

| 0| 0| 0| 0| 1599| 0| 0| 0| 0| 0| 0| 0| 0| 0|

DATOS FALTANTES ACCIDENTES

DATOS FALTANTES ARRESTOS

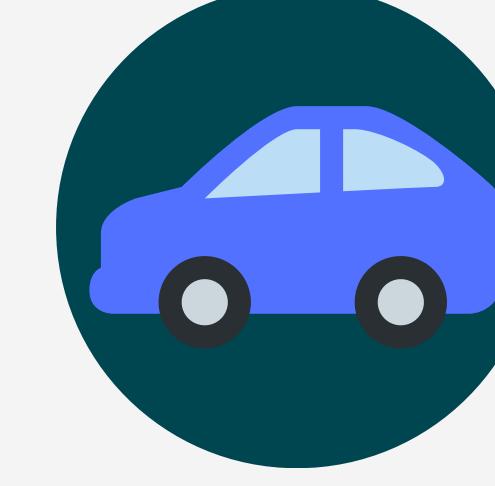
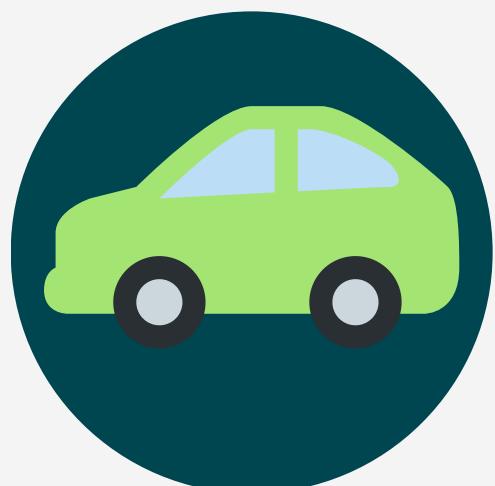
-Grupos por Edad y Género



law_cat_cd
más repetido

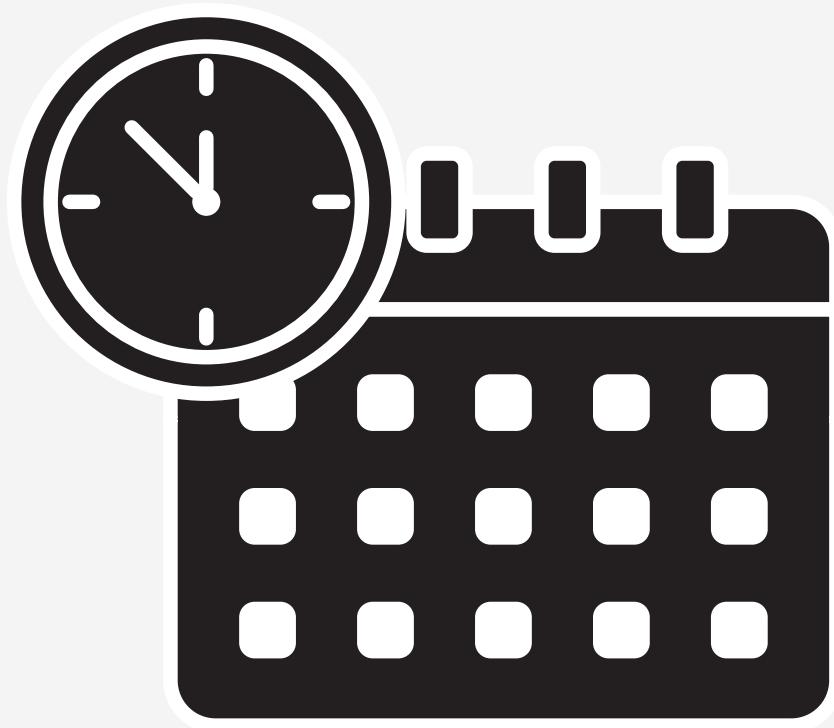
DATOS FALTANTES ACCIDENTES

-Grupos por modelo del vehículo



tipo de vehículo
más repetido

DATOS FALTANTES ARRESTOS

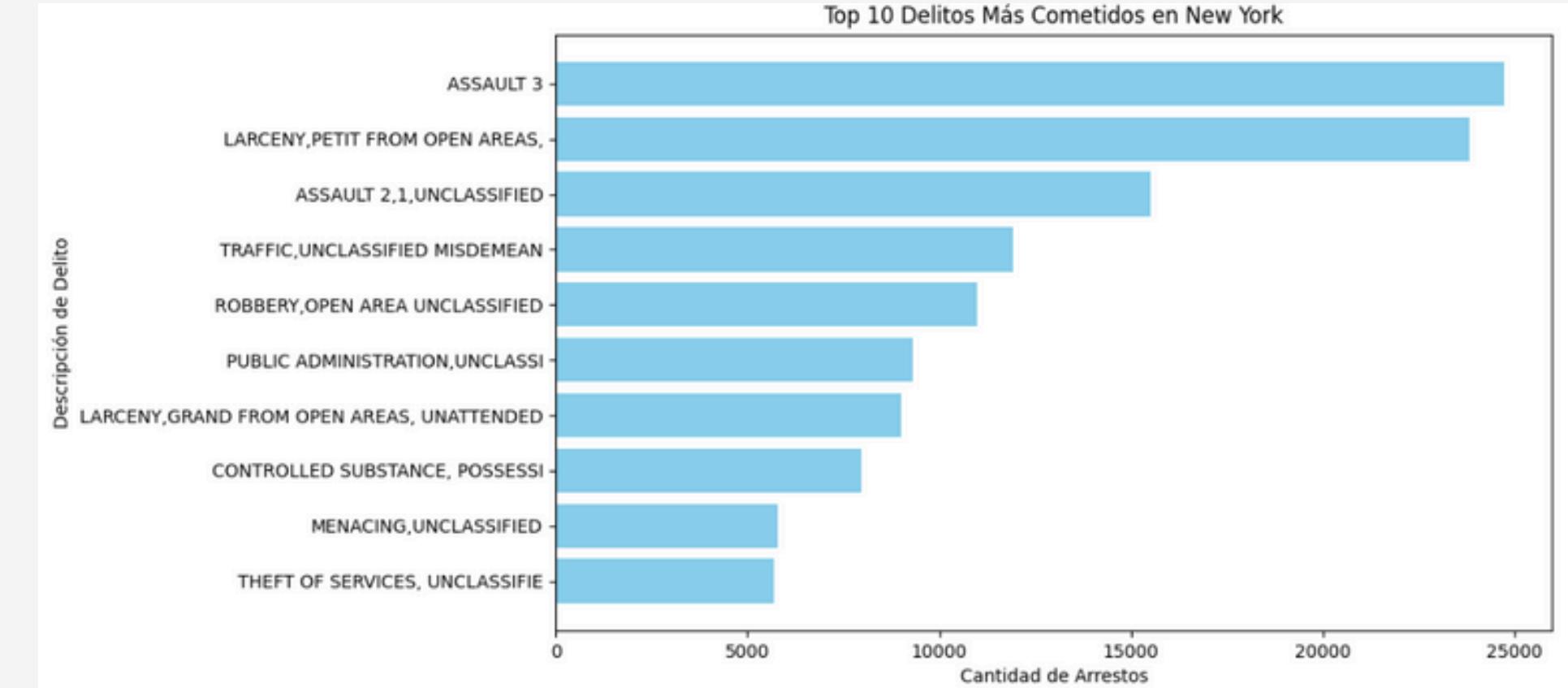
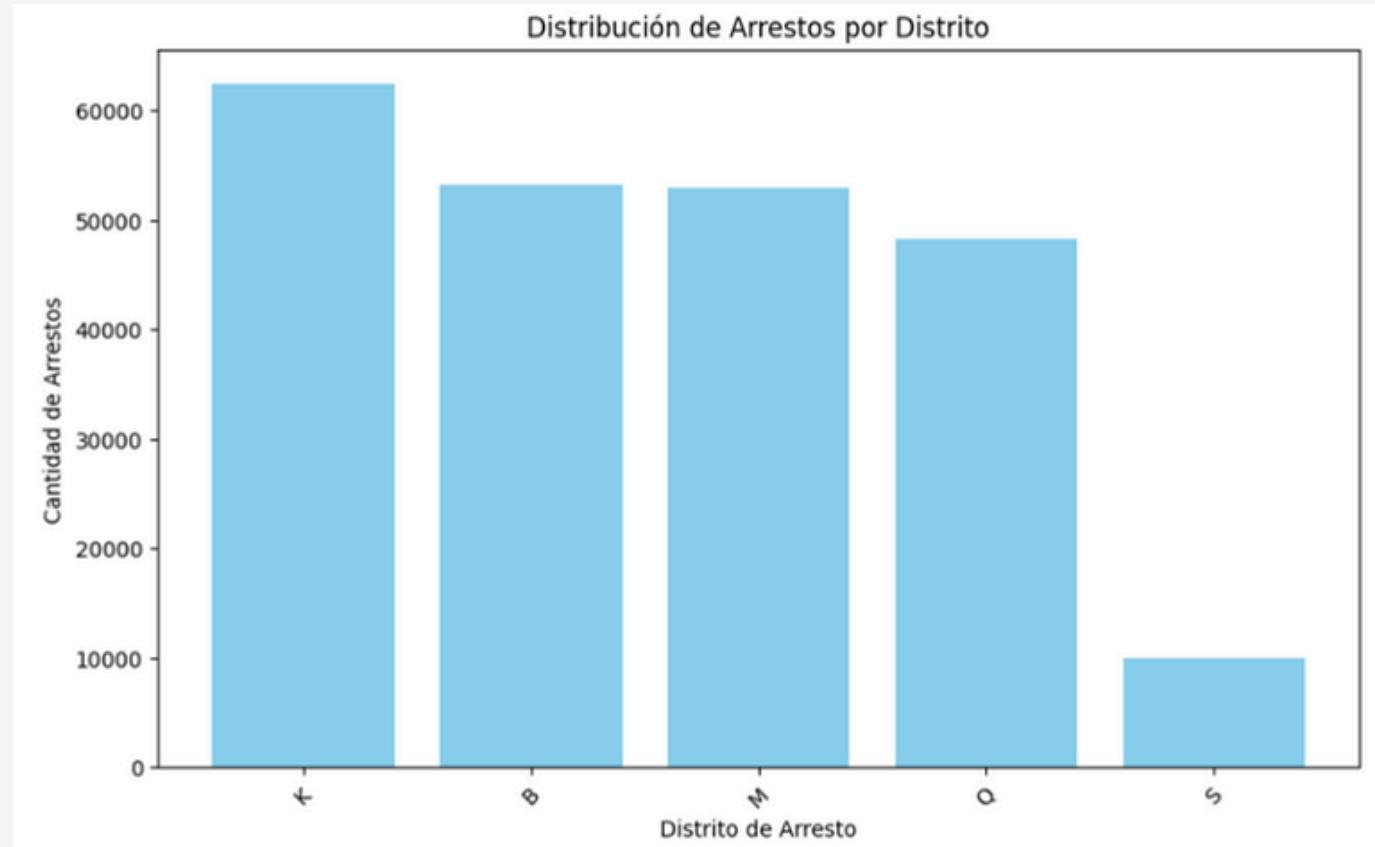


Arrest_Date
Crash_Date

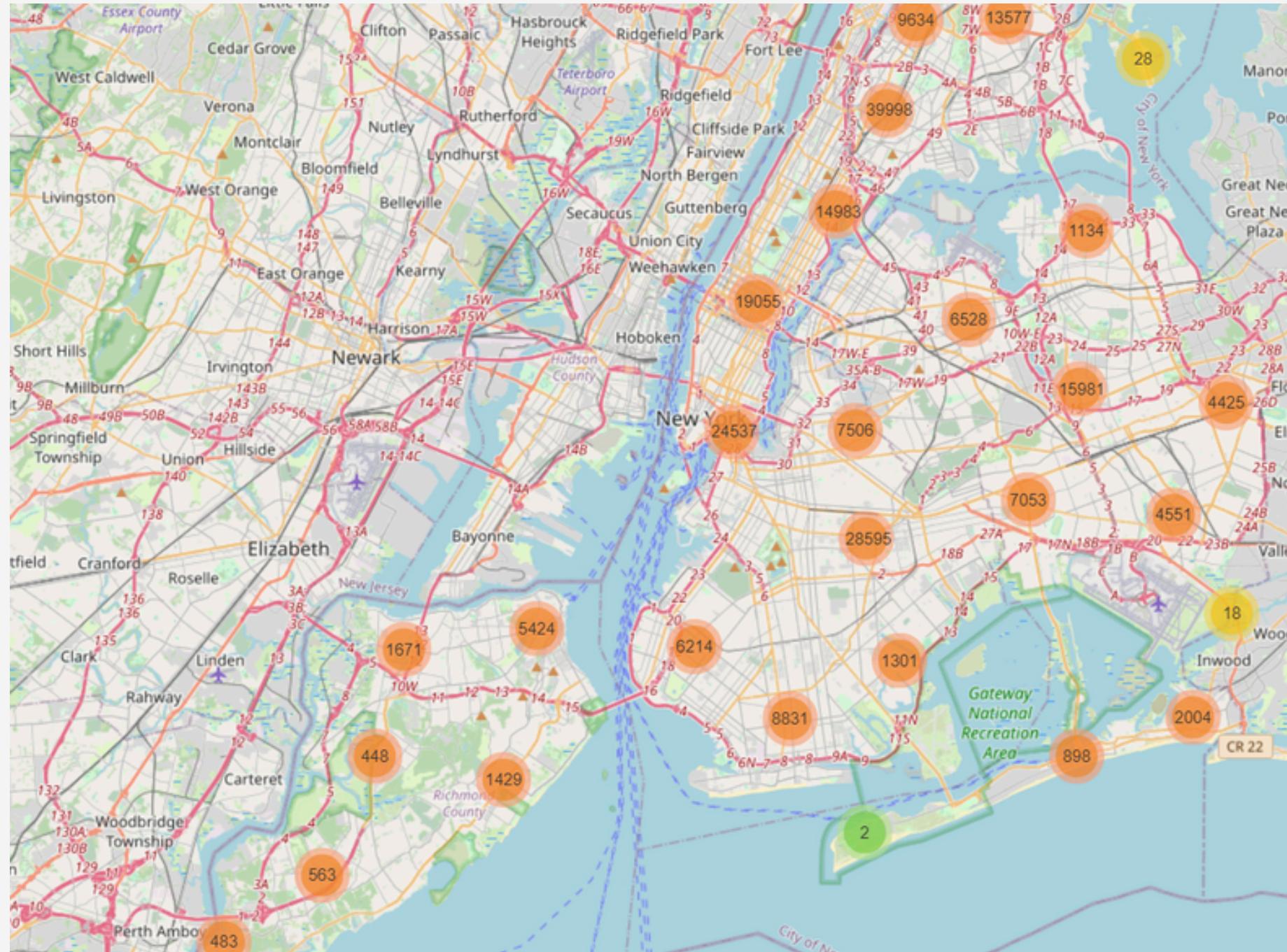
Arrest_Month
Arrest_Year
Crash_Month
Crash_Year

Respuesta a preguntas de negocio

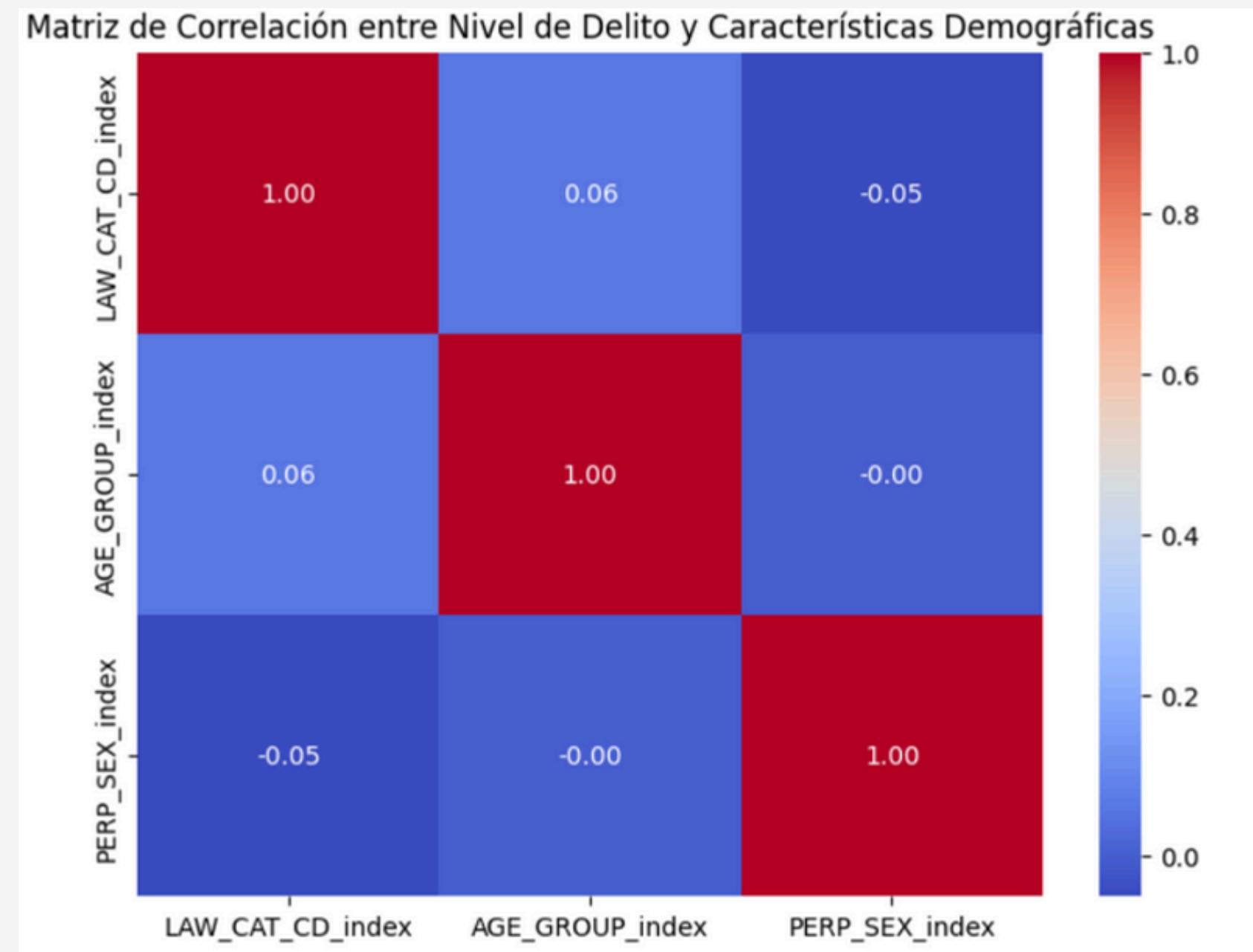
¿Qué tipos de delitos son los más frecuentes y cómo se distribuyen geográficamente?



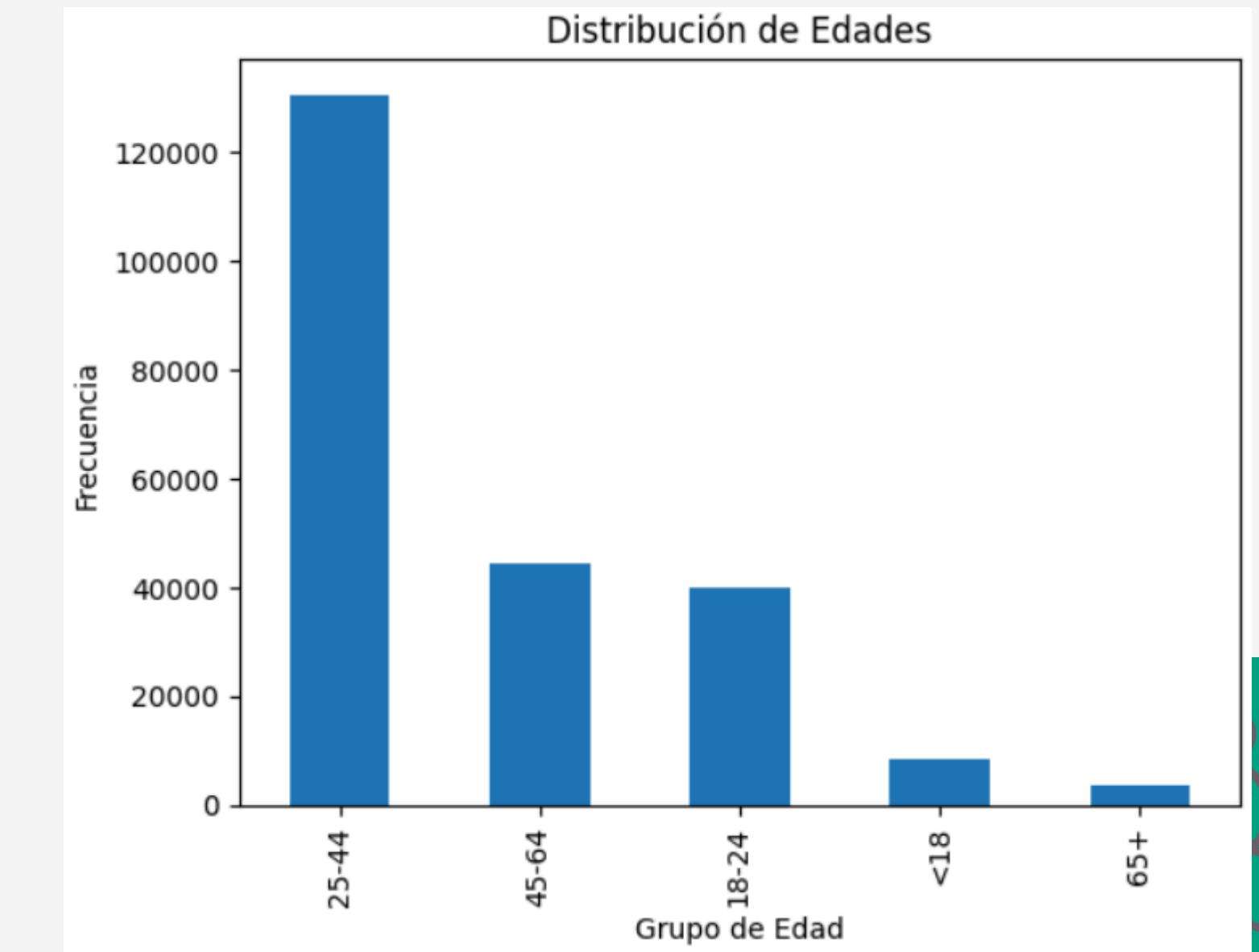
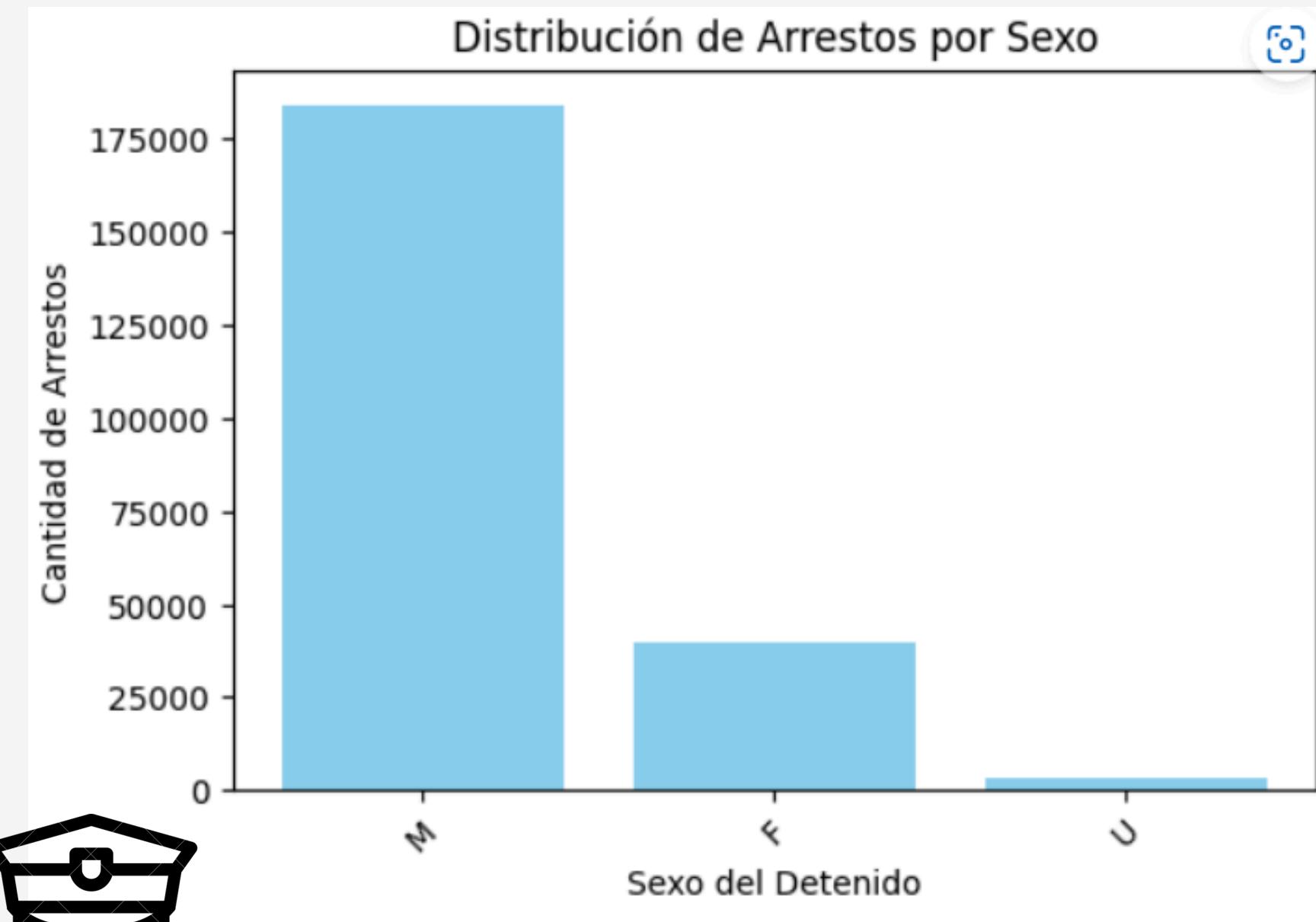
¿Qué tipos de delitos son los más frecuentes y cómo se distribuyen geográficamente?



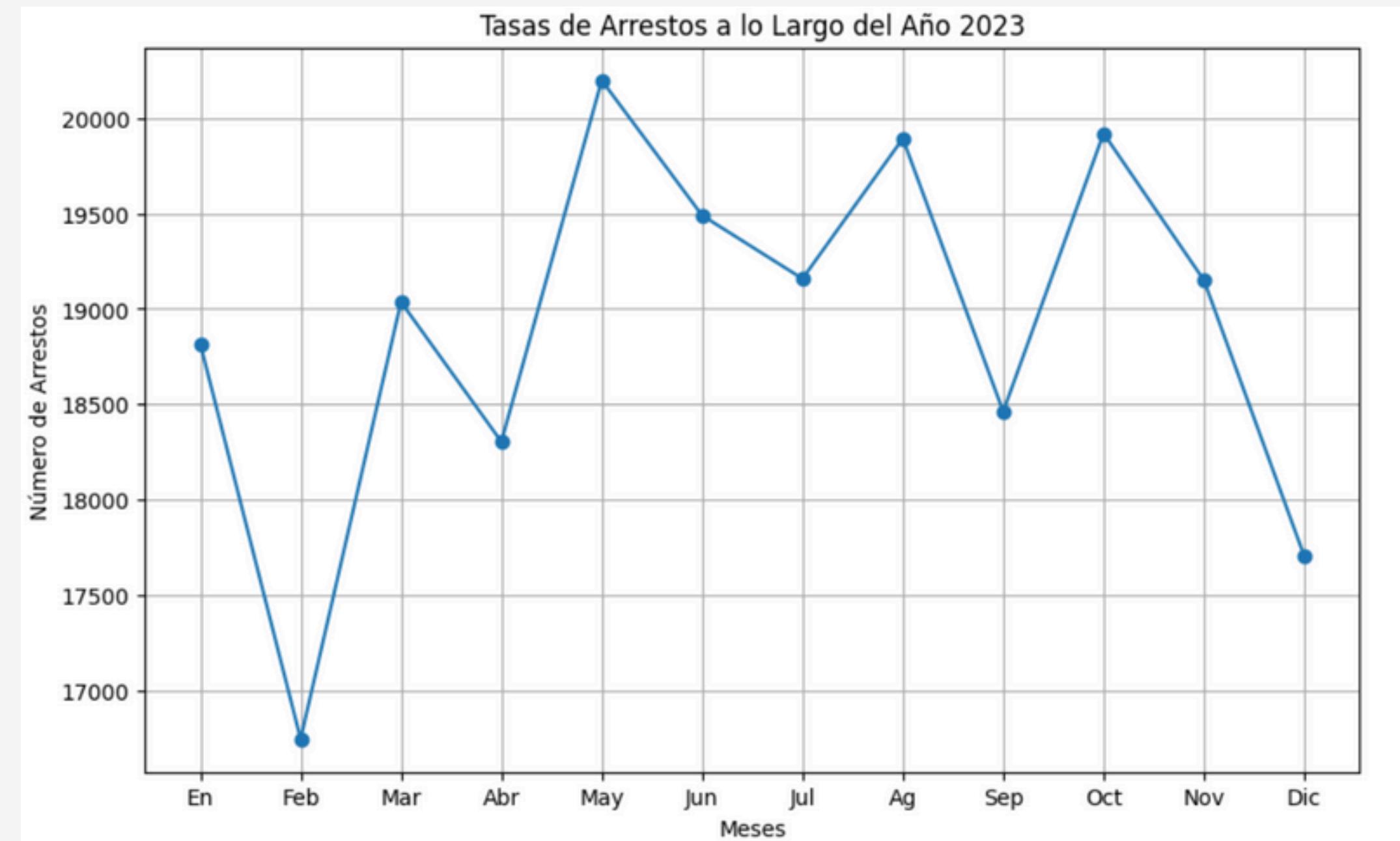
¿Existe alguna relación entre el nivel de delito y las características demográficas de los sospechosos, como su grupo de edad o género ?



¿Existe alguna relación entre el nivel de delito y las características demográficas de los sospechosos, como su grupo de edad o género ?



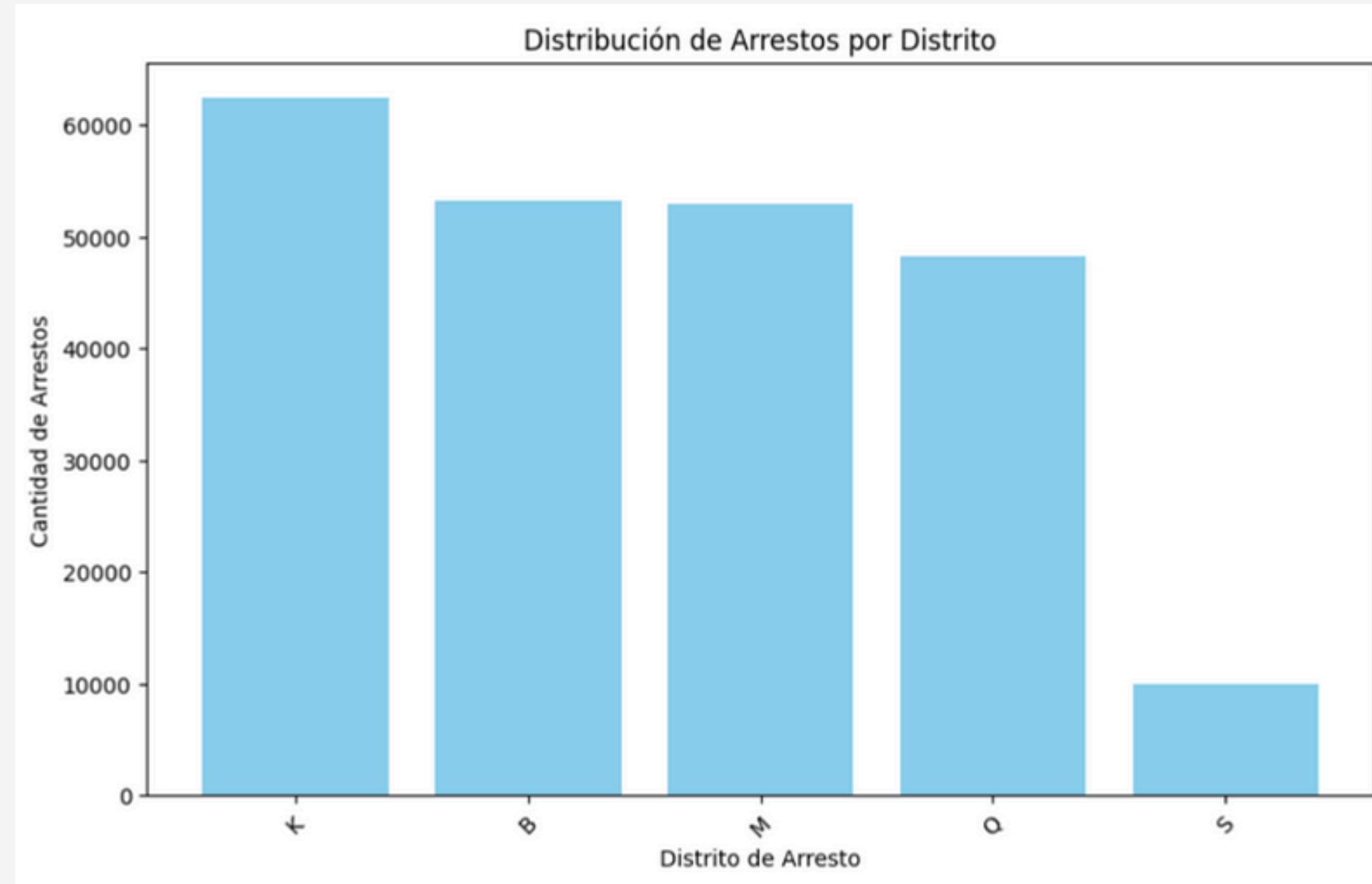
¿Cómo han variado las tasas de arrestos a lo largo del tiempo y si existen patrones temporales o estacionales en los diferentes tipos de delitos?



ARREST_YEAR	ARREST_MONTH	count
2023	1	18817
2023	2	16744
2023	3	19036
2023	4	18303
2023	5	20198
2023	6	19490
2023	7	19158
2023	8	19893
2023	9	18461
2023	10	19920
2023	11	19150
2023	12	17702



¿Qué localidades tienen las mayores tasas de arrestos y cuáles son los tipos de delitos predominantes en esas áreas?



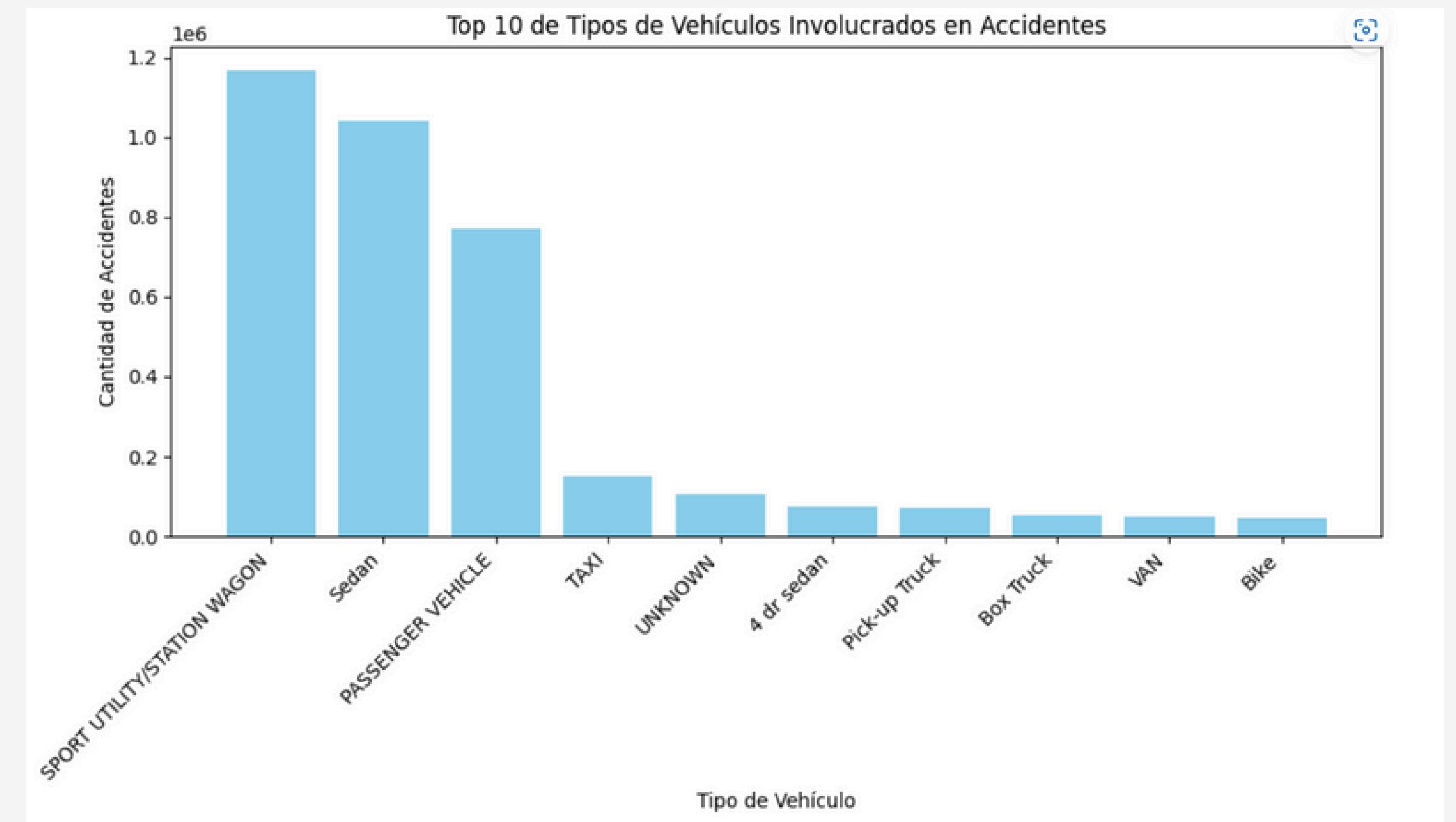
ARREST_BORO	LAW_CAT_CD	count	rank
B	M	30878	1
B	F	21805	2
B	V	187	3
K	M	32149	1
K	F	29189	2
K	V	754	3
M	M	29711	1
M	F	21994	2
M	V	245	3
Q	M	26014	1
Q	F	21369	2
Q	V	118	3
S	M	5729	1
S	F	4264	2
S	V	8	3

delito grave(F)
delito menor(M)
violación(V)

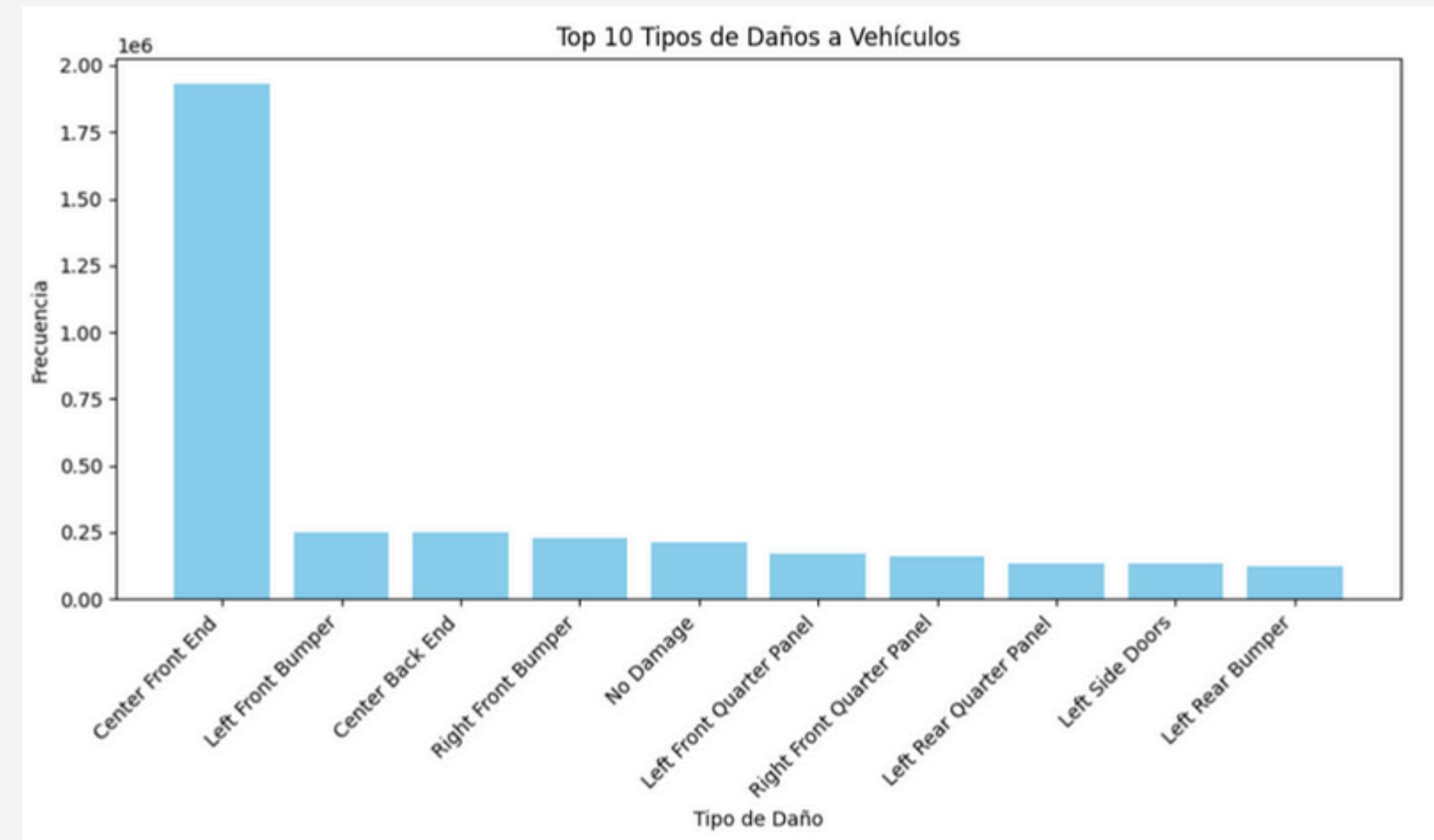
**Brooklyn (K), Bronx (B), Queens (Q),
Manhattan (M), Staten Island (S).**



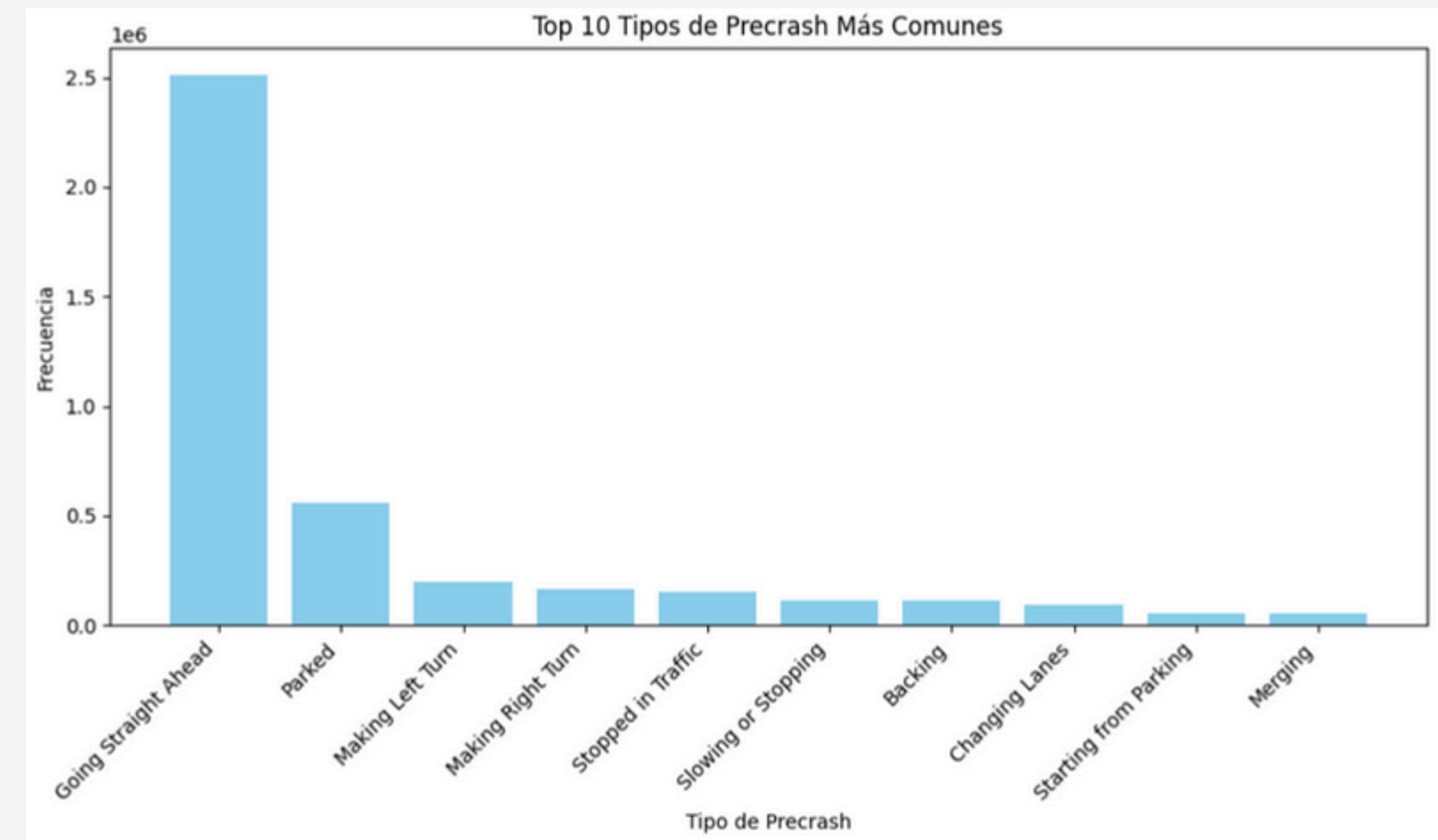
¿Existe un patrón de accidentes relacionado con tipos específicos de vehículos?



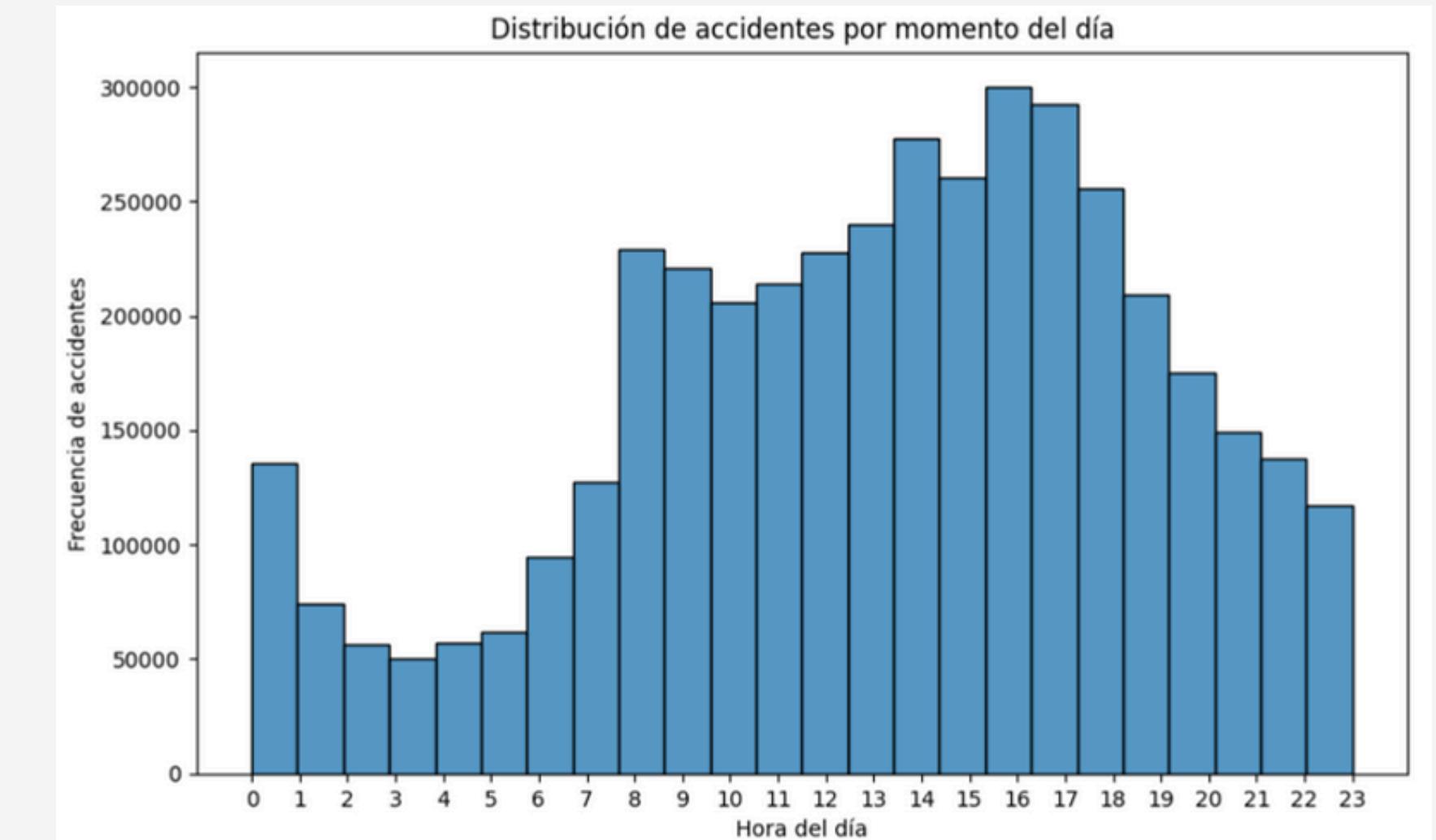
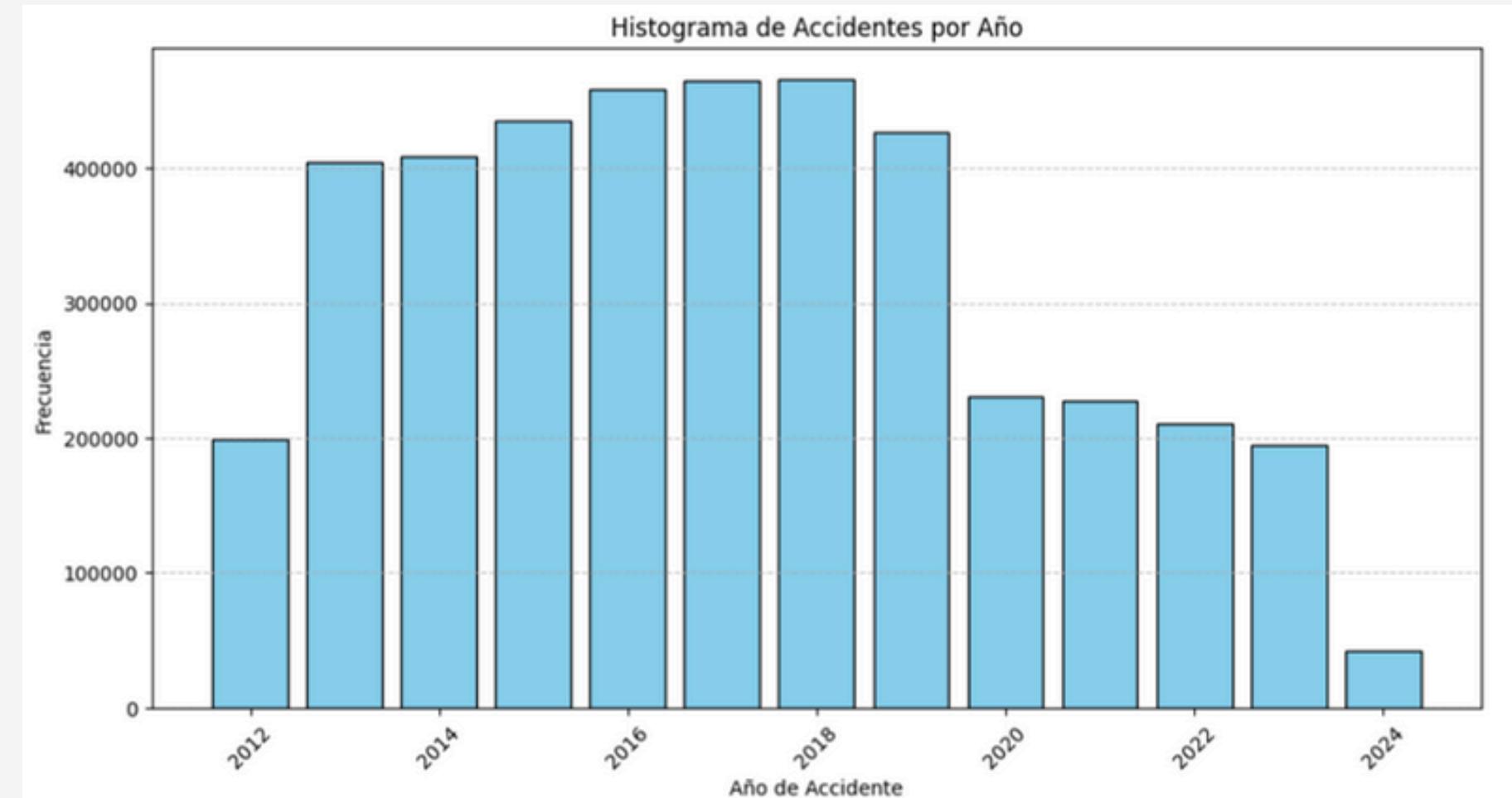
¿Es frecuente que los autos tengan daños en lugares específicos después de un accidente?



¿Existe un patrón en las acciones que realizaban los vehículos justo antes de los accidentes?



¿Los vehículos involucrados en accidentes presentan patrones específicos en las fechas y horas de los choques que podrían ayudar a prevenir futuros incidentes?



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 asignación 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.



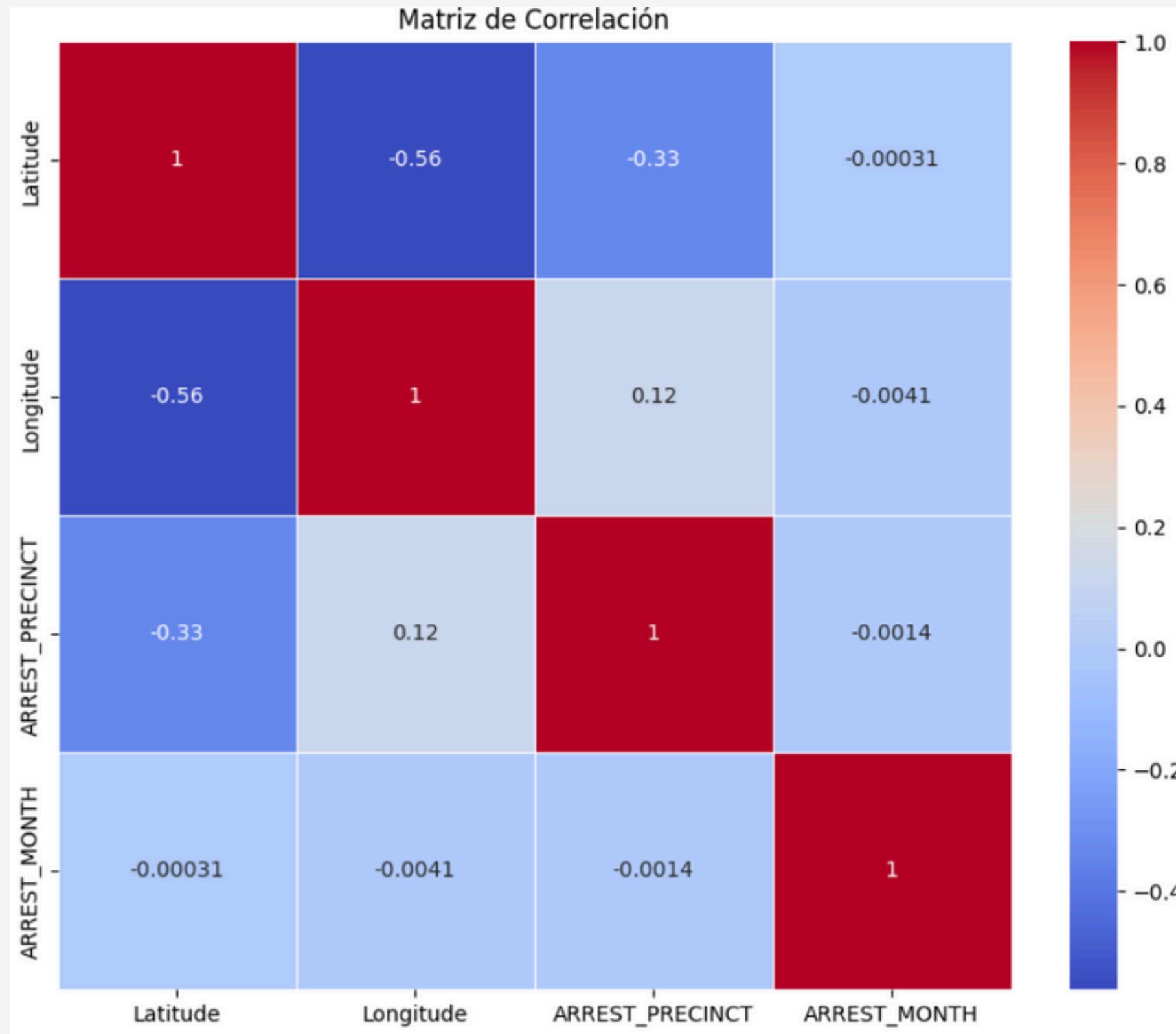
Preparación de datos para modelado

Datos de arrestos del Departamento de Policía de Nueva York



[Volver a la página de agenda](#)

Eliminar características fuertemente correlacionada



Normalización utilizando StandardScaler

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	-0.4533872676644313	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



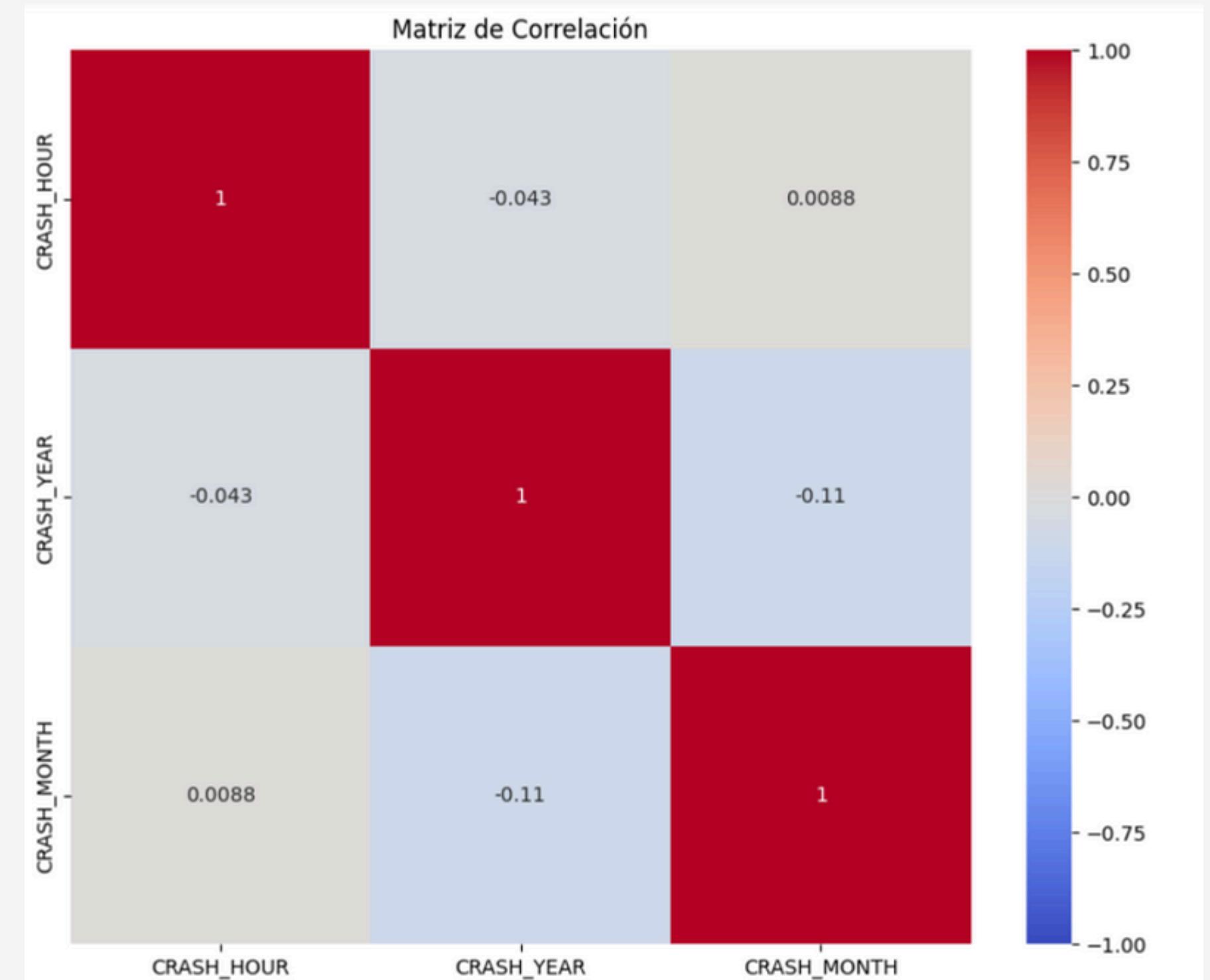
Colisiones de vehículos motorizados - Vehículos



[Volver a la página de agenda](#)

Colisiones de vehículos motorizados - Vehículos

Eliminar características fuertemente correlacionada



Normalización utilizando StandardScaler

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



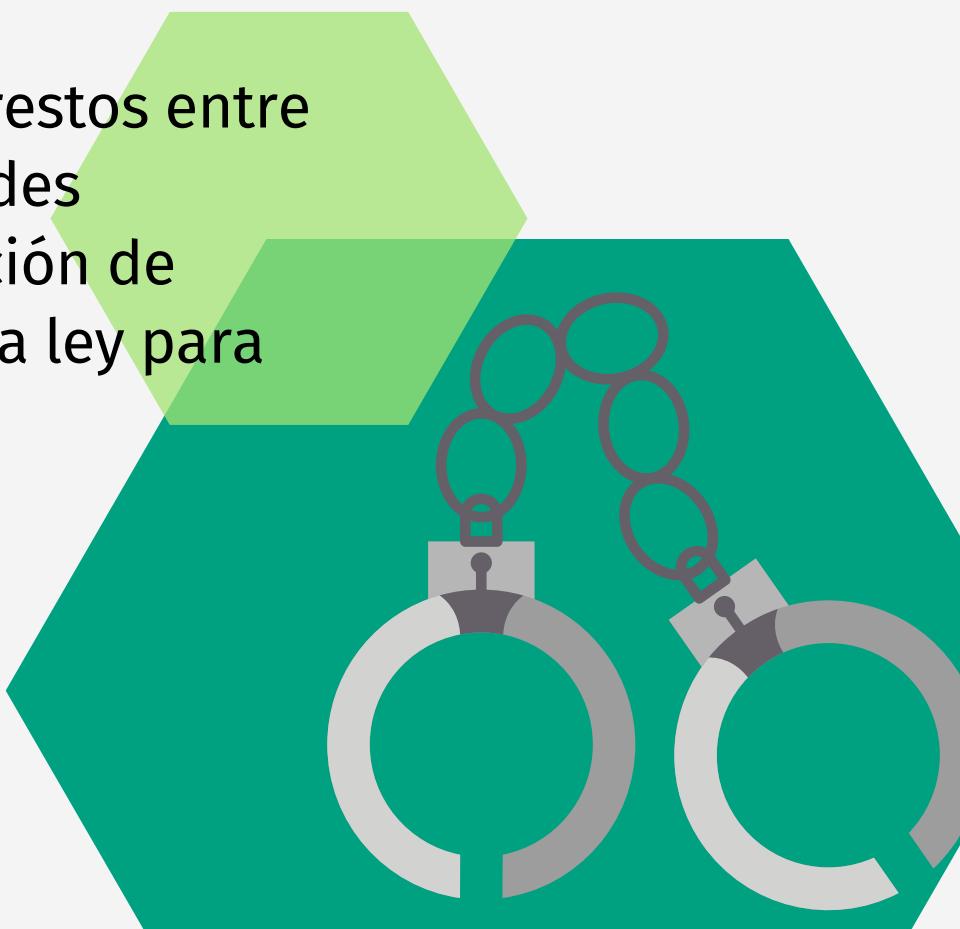
Selección de variables según criterio de negocio

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.



Desarrollo técnicas seleccionadas

Métricas

Implementación de Regresión Logística con PySpark

```
# Indexar la variable categórica objetivo
indexer = StringIndexer(inputCol="PERP_RACE", outputCol="label").fit(df1_scaled)
df_indexed = indexer.transform(df1_scaled)

# Asamblea de features numéricas
feature_cols = ['ARREST_PRECINCT', 'Latitude', 'Longitude', 'ARREST_YEAR', 'ARREST_MONTH']
assembler = VectorAssembler(inputCols=feature_cols, outputCol="features")
df_transformed = assembler.transform(df_indexed)

# Dividir los datos en conjuntos de entrenamiento y prueba
(train_data, test_data) = df_transformed.randomSplit([0.7, 0.3], seed=42)

# Definir y entrenar el modelo de regresión logística con diferentes parámetros
reg_params = [0.01, 0.1, 1.0] # Diferentes parámetros de regularización
results = []
```

- **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'.

Modelo de Regresión Logística

```
# Definir y entrenar el modelo de regresión logística con diferentes parámetros
reg_params = [0.01, 0.1, 1.0] # Diferentes parámetros de regularización
results = []

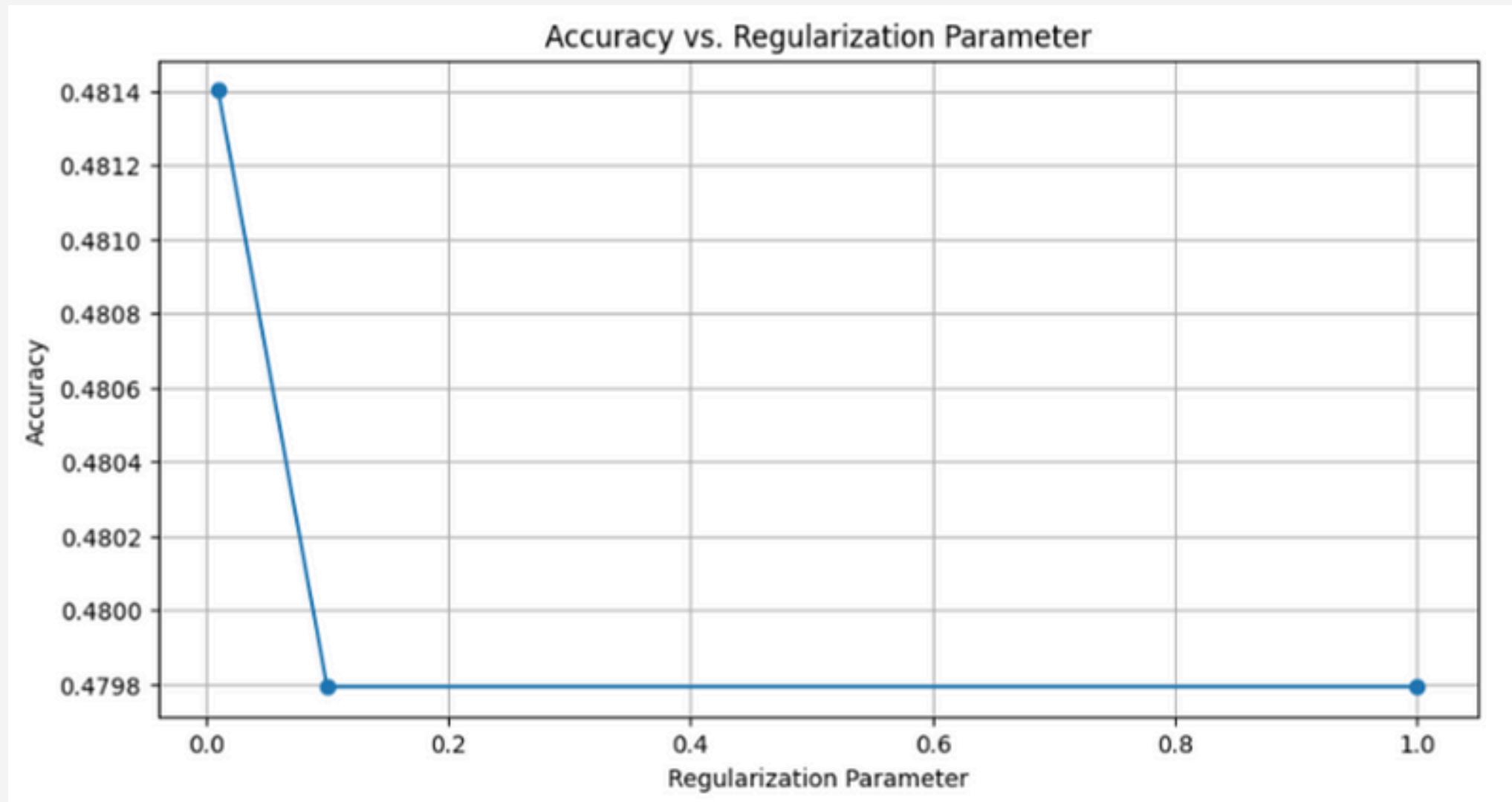
for reg in reg_params:
    lr = LogisticRegression(featuresCol="features", labelCol="label", regParam=reg)
    lr_model = lr.fit(train_data)
    predictions = lr_model.transform(test_data)

    # Evaluar el modelo
    evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="accuracy")
    accuracy = evaluator.evaluate(predictions)
    results.append((reg, accuracy))

    # Matriz de confusión
    rdd = predictions.select(['prediction', 'label']).rdd.map(tuple)
    metrics = MulticlassMetrics(rdd)
    print("Confusion Matrix for regParam = {}".format(reg))
    print(metrics.confusionMatrix().toArray())
```

- **Entrenamiento:** Definición y entrenamiento del modelo con parámetros de regularización [0.01, 0.1, 1.0].
- **Evaluación:** Uso de **MulticlassClassificationEvaluator** para calcular la precisión y presentación de matrices de confusión para cada parámetro.

Implementación de Regresión Logística con PySpark



- 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.
- **Efectividad del Modelo:** La regresión logística, aplicada con PySpark, demuestra ser una herramienta eficaz para la clasificación en grandes volúmenes de datos, como se refleja en la precisión obtenida.

[Volver a la página de agenda](#)

Análisis de Clustering con PySpark y Visualización PCA

```
# Asegurarse de que los datos están transformados
df_transformed = assembler.transform(df1_scaled)

# Prueba de diferentes números de clústeres
cluster_counts = [2, 3, 5, 7]
silhouettes = []

for k in cluster_counts:
    kmeans = KMeans(featuresCol="features", k=k, seed=42)
    model = kmeans.fit(df_transformed)
    predictions = model.transform(df_transformed)

    # Evaluar el modelo
    evaluator = ClusteringEvaluator()
    silhouette = evaluator.evaluate(predictions)
    silhouettes.append(silhouette)
    print("Silhouette score for k={} is {}".format(k, silhouette))

# Realizar PCA para reducir a 2 dimensiones
pca = PCA(k=2, inputCol="features", outputCol="pcaFeatures")
model_pca = pca.fit(predictions) # Asegurarse de usar 'predictions'
result_pca = model_pca.transform(predictions)

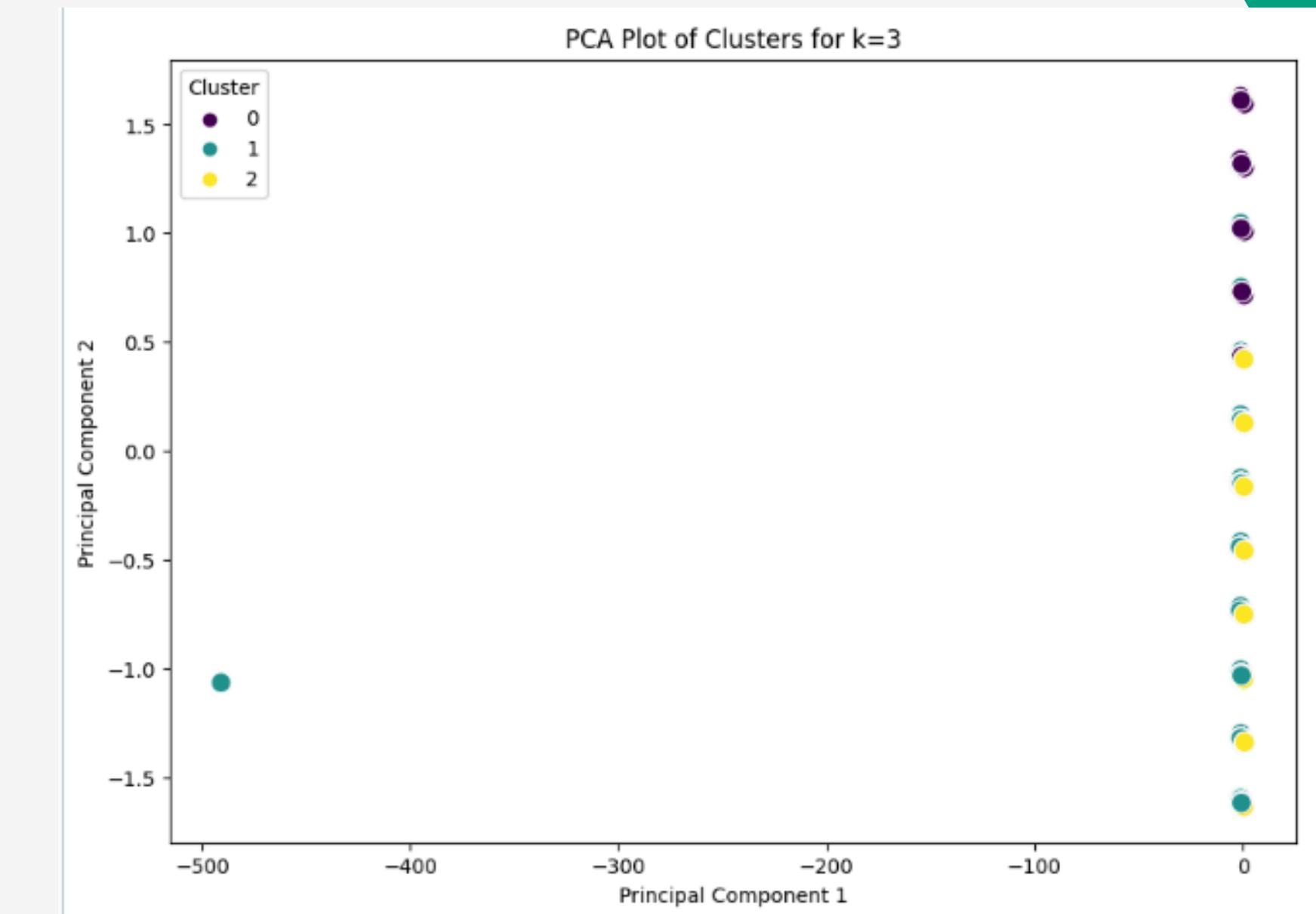
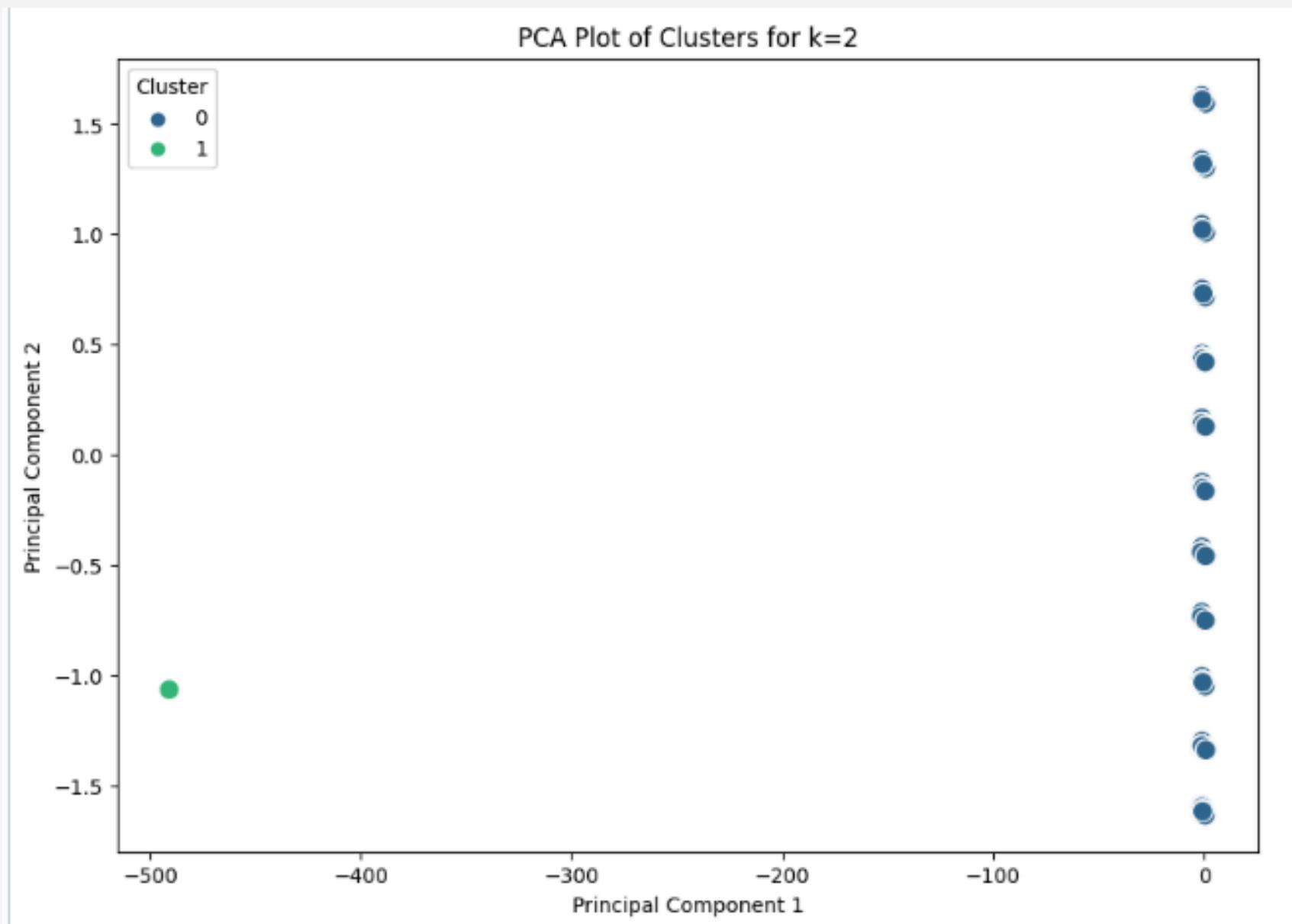
# Extraer las coordenadas PCA y las etiquetas de clúster para graficar
pca_features = result_pca.select("pcaFeatures", "prediction").collect()
x_pca = [feat.pcaFeatures[0] for feat in pca_features]
y_pca = [feat.pcaFeatures[1] for feat in pca_features]
cluster_labels = [feat.prediction for feat in pca_features]
```

- Transformación de Datos: Utilización de VectorAssembler para preparar el dataframe df1_scaled con las características adecuadas.

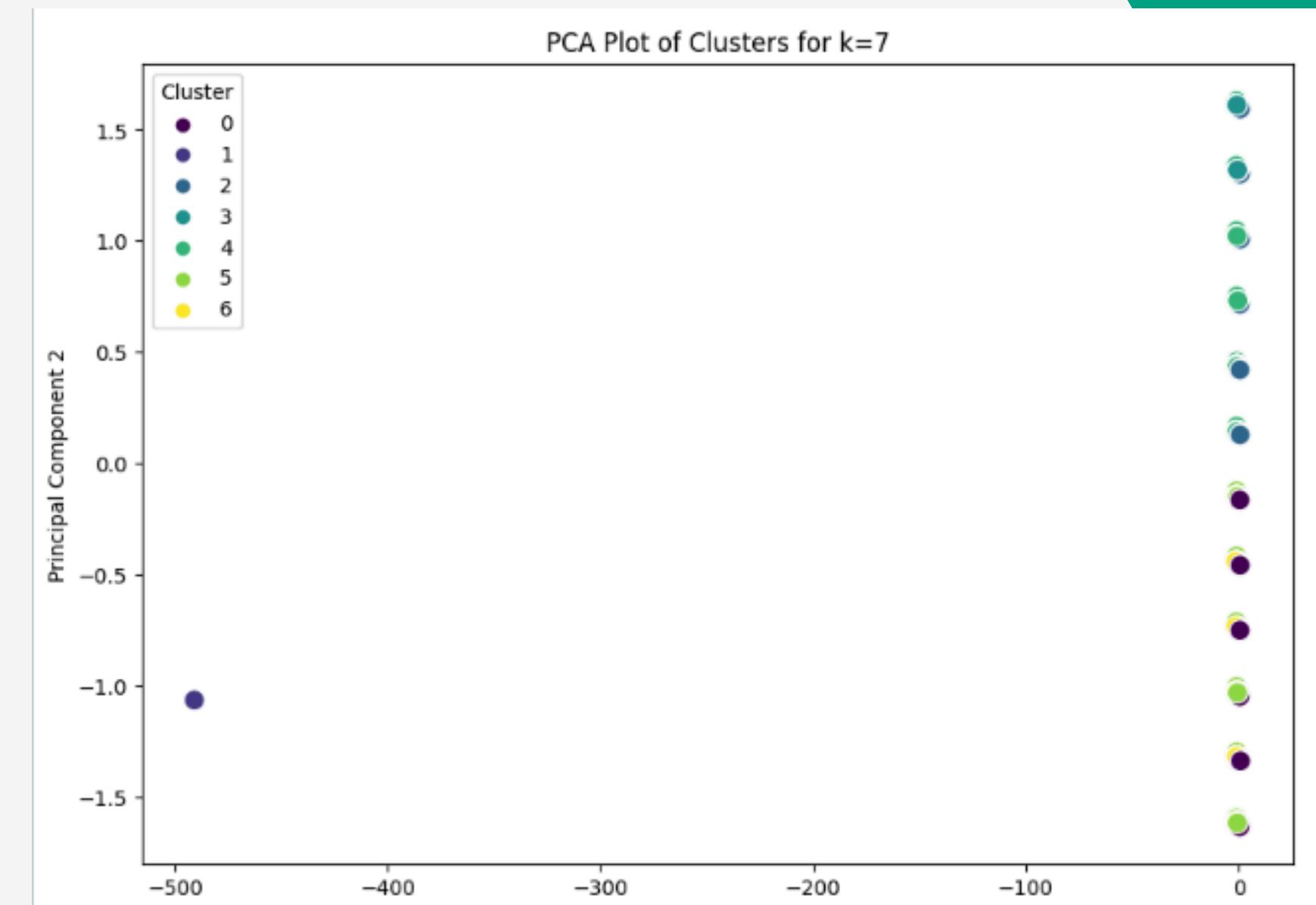
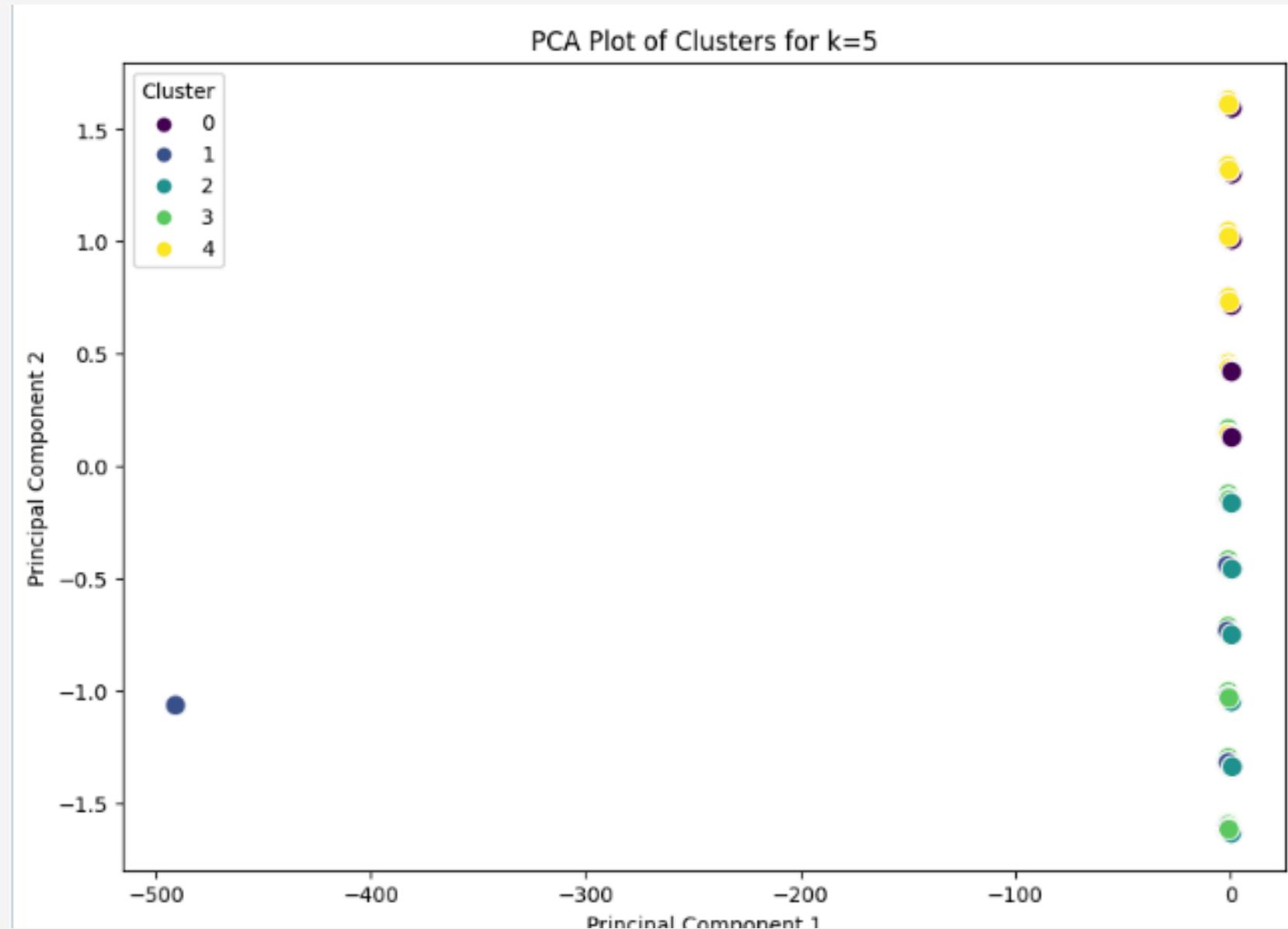
Análisis de Clustering

- **Pruebas con Diferentes Clústeres:** Experimentación con distintos números de clústeres (2, 3, 5, 7) para identificar la agrupación óptima.
- **Evaluación de Modelos:** Uso de ClusteringEvaluator para calcular el score de silueta, que mide la calidad de los clústeres formados.

Análisis de Clustering con PySpark y Visualización PCA



Análisis de Clustering con PySpark y Visualización PCA



La gráfica de PCA muestra la distribución de siete clústeres en dos dimensiones principales, destacando una separación notable entre algunos grupos, como el clúster en azul oscuro, que está visiblemente distanciado de los demás, indicando características únicas en sus observaciones. Otros clústeres aparecen más cercanos entre sí, sugiriendo similitudes en algunas dimensiones pese a estar clasificados en grupos distintos. Este tipo de visualización es útil para evaluar la efectividad del clustering en la reducción de la dimensionalidad y para explorar las relaciones intrínsecas entre los grupos, pudiendo indicar si ajustes en el número de clústeres o en la configuración del modelo podrían mejorar la segmentación.

BIBLIOGRAFIA

<https://kinsta.com/es/base-de-conocimiento/que-es-web-scraping/#:~:text=El%20web%20scraping%20se%20refiere,precios%20de%20varias%20tiendas%20online.>

<https://www.kyoceradocumentsolutions.es/es/smarter-workspaces/business-challenges/the-cloud/procesamiento-datos-grandes-servidores.html>

<https://spark.apache.org/docs/3.3.1/api/python/index.html>

<https://catalog.data.gov/dataset/nypd-arrest-data-year-to-date>

GRACIAS