PONTIFICIA UNIVERSIDAD JAVERIANA

Proyecto – Entrega #1

Anamaria Leguizamon- Diego Herrera -Sofia Galindo

Procesamiento de Datos a Gran Escala

Ciencia de Datos

Primer Semestre 2024
10/04/2024

Entendimiento del negocio:

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.

1. **Objetivo General:**

El objetivo general de este proyecto es desarrollar una solución basada en Big Data que permita mejorar la cantidad de arrestos y la cantidad de accidentes viales en Nueva York, utilizando la metodología CRISP-DM.

2. **Objetivos específicos:**

1. Identificar patrones en las zonas con mayor número de accidentes.
2. Identificar patrones en las zonas con mayor número de delitos.
3. Analizar el impacto de otras variables en los indicadores de interés.
4. Detectar y prevenir las principales causas de delitos y accidentes viales.

Selección de los datos a utilizar:

A continuación se presenta una lista de los conjuntos de datos que vamos a usar durante el proyecto con su respectiva justificación:

Datos de arrestos del Departamento de Policía de Nueva York hasta la fecha:

Este conjunto de datos es clave para analizar los patrones de arrestos en la ciudad de Nueva York, lo que nos permite identificar áreas y momentos de alta criminalidad. Comprender estos patrones es importante para implementar intervenciones específicas para reducir las tasas de criminalidad de manera efectiva.

Colisiones de vehículos motorizados - Vehículos:

La seguridad vial es una preocupación importante en las zonas urbanas. Al analizar los datos sobre accidentes de tránsito, podemos identificar causas comunes y ubicaciones de

colisiones, lo cual es esencial para implementar medidas de seguridad y reducir las muertes y lesiones relacionadas con el tránsito.

Colección y Descripción de Datos

En esta sección, se presenta una descripción detallada de los datos utilizados en el proyecto, incluyendo los tipos de datos presentes y la comprensión del significado de cada atributo.

1. **Carga de Datos en Databricks**

Los datos utilizados en este proyecto fueron cargados en el ambiente de Databricks utilizando la siguiente configuración:

- Fuente de datos: Archivo CSV
- Opciones de carga: Formato: CSV; Inferencia de esquema: Activada, Encabezado: Activado; Separador de campos: ",".

2. **Tipos de Datos**

Se utilizaron dos conjuntos de datos diferentes en este proyecto, cada uno con su propia estructura y tipos de datos. En el conjunto de datos de "Arrestos del Departamento de Policía de Nueva York hasta la fecha", predominan los datos de tipo entero, fecha y cadena de texto, que representan claves de arresto, fechas de arresto, códigos de delitos, descripciones de delitos, etc. Por último, en el conjunto de datos "Colisiones de vehículos motorizados - Vehículos", se encuentran una variedad de tipos de datos, como enteros, fechas, marcas y modelos de vehículos, así como descripciones de los daños y factores contribuyentes a los accidentes. Cada conjunto de datos proporciona una visión única y complementaria para el análisis realizado en este proyecto.

3. **Comprensión del Significado de Cada Atributo**

Datos de arrestos del Departamento de Policía de Nueva York hasta la fecha:

ARREST_KEY: Clave única que identifica cada arresto.
ARREST_DATE: Fecha en la que ocurrió el arresto.
PD_CD: Código del departamento de policía asociado al delito.
PD_DESC: Descripción del delito.
KY_CD: Código de clasificación del delito.
OFNS_DESC: Descripción de la clasificación del delito.
LAW_CODE: Código de la ley asociada al delito.
LAW_CAT_CD: Categoría legal del delito.
ARREST_BORO: Distrito donde se realizó el arresto.
ARREST_PRECINCT: Número de la comisaría donde se realizó el arresto.
JURISDICTION_CODE: Código de jurisdicción.
AGE_GROUP: Grupo de edad del individuo arrestado.
PERP_SEX: Sexo del individuo arrestado.
PERP_RACE: Raza del individuo arrestado.
X_COORD_CD: Coordenada X de ubicación del arresto.
Y_COORD_CD: Coordenada Y de ubicación del arresto.
Latitude: Latitud de la ubicación del arresto.

Longitude: Longitud de la ubicación del arresto.
New Georeferenced Column: Columna adicional de georreferencia (puede contener información adicional sobre la ubicación).

Colisiones de vehículos motorizados - Vehículos:

UNIQUE_ID: Identificación única del registro generado por el sistema. Clave primaria. (Número)
COLLISION_ID: Código de identificación de colisión. Clave foránea, coincide con unique_id de la tabla de Colisiones. (Número)
CRASH_DATE: Fecha de ocurrencia de la colisión. (Fecha y Hora)
CRASH_TIME: Hora de ocurrencia de la colisión. (Texto)
VEHICLE_ID: Código de identificación del vehículo asignado por el sistema. (Texto)
STATE_REGISTRATION: Estado donde está registrado el vehículo. (Texto)
VEHICLE_TYPE: Tipo de vehículo según la categoría de vehículo seleccionada . (Texto)
VEHICLE_MAKE: Marca del vehículo. (Texto)
VEHICLE_MODEL: Modelo del vehículo. (Texto)
VEHICLE_YEAR: Año de fabricación del vehículo. (Texto)
TRAVEL_DIRECTION: Dirección en la que se desplazaba el vehículo. (Texto)
VEHICLE_OCCUPANTS: Número de ocupantes del vehículo. (Número)
DRIVER_SEX: Género del conductor. (Texto)
DRIVER_LICENSE_STATUS: Estado de la licencia del conductor, permiso, sin licencia. (Texto)
DRIVER_LICENSE_JURISDICTION: Estado donde se emitió la licencia de conducir. (Texto)
PRE_CRASH: Acción previa al choque: ir en línea recta, girar a la derecha, pasar, retroceder, etc. (Texto)
POINT_OF_IMPACT: Ubicación en el vehículo del punto inicial de impacto (es decir, lado del conductor, parte trasera del lado del pasajero, etc.). (Texto)
VEHICLE_DAMAGE: Ubicación en el vehículo donde ocurrió la mayor parte del daño. (Texto)
VEHICLE_DAMAGE_1: Ubicaciones de daños adicionales en el vehículo. (Texto)
VEHICLE_DAMAGE_2: Ubicaciones de daños adicionales en el vehículo. (Texto)
VEHICLE_DAMAGE_3: Ubicaciones de daños adicionales en el vehículo. (Texto)
PUBLIC_PROPERTY_DAMAGE: Propiedad pública dañada (Sí o No). (Texto)
PUBLIC_PROPERTY_DAMAGE_TYPE: Tipo de propiedad pública dañada (por ejemplo, señal, cerca, poste de luz, etc.). (Texto)
CONTRIBUTING_FACTOR_1: Factores que contribuyen a la colisión para el vehículo designado. (Texto)
CONTRIBUTING_FACTOR_2: Factores que contribuyen a la colisión para el vehículo designado. (Texto)

- **Descripción general del contenido de los conjuntos de datos**

El conjunto de datos proviene del Departamento de Policía de Nueva York (NYPD) y fue creado el 5 de junio de 2018, con actualizaciones de datos y metadatos más recientes el 18 de enero de 2024. Cubre los arrestos realizados por el NYPD en los cinco distritos de la ciudad de Nueva York y consta de 227,000 registros individuales, donde cada fila representa un arresto con 19 columnas que contienen detalles como

clave única, fecha, código del distrito policial, descripción del delito, códigos legales, ubicación con coordenadas geográficas y datos demográficos del arrestado (edad, sexo, raza). Estos datos en formato CSV permiten análisis en profundidad de patrones delictivos, asignación estratégica de recursos policiales, estudios sobre posibles sesgos y desigualdades en el sistema de justicia penal, así como investigaciones académicas en criminología, sociología y disciplinas relacionadas. Sin embargo, es importante considerar que pueden existir datos faltantes o incompletos, y se deben tomar precauciones para proteger la privacidad de los individuos y evitar interpretaciones o usos indebidos de la información.

Este conjunto de datos proviene del Departamento de Policía de Nueva York (NYPD) y contiene registros de vehículos motorizados involucrados en choques ocurridos en la ciudad de Nueva York. Fue creado originalmente el 30 de julio de 2019, con la última actualización de datos el 8 de abril de 2024 y la última actualización de metadatos el 17 de septiembre de 2021. El conjunto de datos se proporciona en formato CSV y consta de 4,17 millones de filas, donde cada fila representa un vehículo involucrado en un choque, con 25 columnas que incluyen un identificador único, detalles del choque como fecha, hora y ubicación, información del vehículo como tipo, marca, modelo, año, dirección de viaje y número de ocupantes, datos del conductor como género, estado de la licencia y jurisdicción, información sobre el impacto como ubicación del punto de impacto y daños, si hubo daños a la propiedad pública y de qué tipo, y factores contribuyentes al choque.

Exploración de los datos:

Datos de arrestos del Departamento de Policía de Nueva York hasta la fecha.

librerías utilizadas:

```
import pyspark #contiene todas las funciones principales de PySpark
from pyspark import SparkContext #RDD
from pyspark.sql import SQLContext, Row #funcionalidad para trabajar datos estructurados y row para representar una fila.
from pyspark.sql.functions import col #col para representar una columna.
from pyspark.sql.types import IntegerType, FloatType #para especificar los datos de las columnas.
import pandas as pd #biblioteca para el análisis de datos.
import matplotlib.pyplot as plt
import pyspark.sql.functions as F
import random
import seaborn as sns


sesion de Pyspark
sc= SparkContext.getOrCreate() #se crea el contexto spark para interactuar con sus funciones.
sql_sc = SQLContext(sc) #contexto sql con el que se puede ejecutar
```

```
consultas sql sobre los datos directamente.
sc #se imprime para verificar que se haya creado correctamente.
```

Para empezar se carga el archivo CSV (datos de arrestos del Departamento de Policía de Nueva York hasta la fecha) en el data frame df1, se utiliza las funciones:

```
display (df1)- para imprimir el dataframe
print (df1.count()) - saber el número total de filas
```

Resultado:

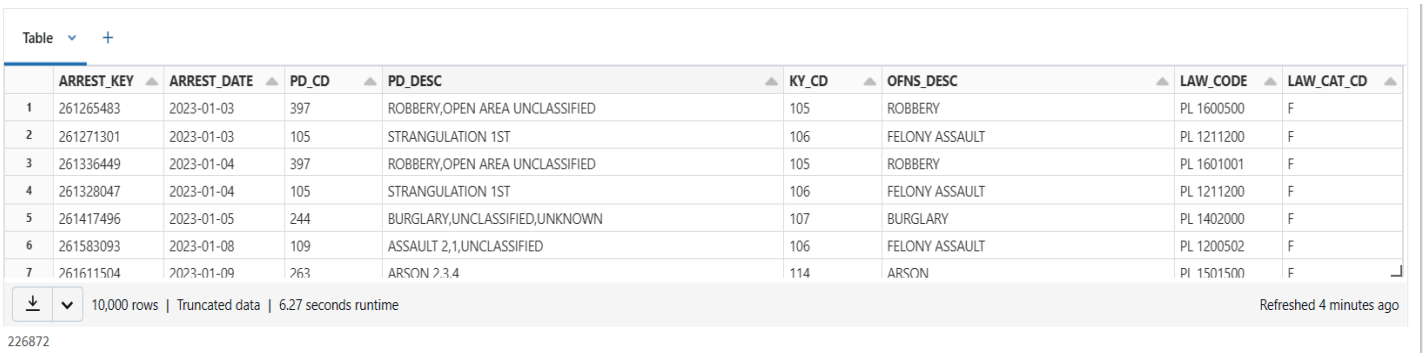


Table +

	ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE	LAW_CAT_CD
1	261265483	2023-01-03	397	ROBBERY,OPEN AREA UNCLASSIFIED	105	ROBBERY	PL 1600500	F
2	261271301	2023-01-03	105	STRANGULATION 1ST	106	FELONY ASSAULT	PL 1211200	F
3	261336449	2023-01-04	397	ROBBERY,OPEN AREA UNCLASSIFIED	105	ROBBERY	PL 1601001	F
4	261328047	2023-01-04	105	STRANGULATION 1ST	106	FELONY ASSAULT	PL 1211200	F
5	261417496	2023-01-05	244	BURGLARY,UNCLASSIFIED,UNKNOWN	107	BURGLARY	PL 1402000	F
6	261583093	2023-01-08	109	ASSAULT 2,1,UNCLASSIFIED	106	FELONY ASSAULT	PL 1200502	F
7	261611504	2023-01-09	263	ARSON 2,3,4	114	ARSON	PL 1501500	F

10,000 rows | Truncated data | 6.27 seconds runtime — Refreshed 4 minutes ago

226872

utilizamos la función: `df1.printSchema()`

```
root
 |-- ARREST_KEY: integer (nullable = true)
 |-- ARREST_DATE: date (nullable = true)
 |-- PD_CD: integer (nullable = true)
 |-- PD_DESC: string (nullable = true)
 |-- KY_CD: integer (nullable = true)
 |-- OFNS_DESC: string (nullable = true)
 |-- LAW_CODE: string (nullable = true)
 |-- LAW_CAT_CD: string (nullable = true)
 |-- ARREST_BORO: string (nullable = true)
 |-- ARREST_PRECINCT: integer (nullable = true)
 |-- JURISDICTION_CODE: integer (nullable = true)
 |-- AGE_GROUP: string (nullable = true)
 |-- PERP_SEX: string (nullable = true)
 |-- PERP_RACE: string (nullable = true)
 |-- X_COORD_CD: integer (nullable = true)
 |-- Y_COORD_CD: integer (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- New Georeferenced Column: string (nullable = true)
```

Se utiliza la función `display(df1.describe()`)genera un resumen estadístico del Data Frame df1 y lo muestra en formato tabular. Este resumen incluye estadísticas descriptivas como la cuenta (count), media (mean), desviación estándar (stddev), mínimo (min), máximo (max) y algunos percentiles (25%, 50%, 75%) para todas las columnas numéricas del DataFrame.

Resultado:

Table +

	summary	ARREST_KEY	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE	LAW_CAT_CD	ARREST_BORO	A
1	count	226872	226870	226872	226855	226872	226872	225273	226872	22
2	mean	2.7064792484001111E8	424.7544011989245	null	249.3451323532653	null	null	9.0	null	63
3	stddev	5304010.298147567	274.4753806048603	null	147.6867326476052	null	null	0.0	null	34
4	min	261180920	1	(null)	101	(null)	(null)	(null)	B	1
5	max	279779734	997	WEAPONS,MFR,TRANSPORT,ETC.	995	VEHICLE AND TRAFFIC LAWS	VTL21300A5	V	S	12

Se hace un conteo de valores únicos para variables categóricas

```
df1.select("PERP_SEX",       "PERP_RACE",       "PD_DESC").groupBy("PERP_SEX",
"PERP_RACE", "PD_DESC").count().show()
```

Resultado:

```
--------+--------------------+--------------------+-----+
PERP_SEX|           PERP_RACE|             PD_DESC|count|
--------+--------------------+--------------------+-----+
       F|      WHITE HISPANIC|NY STATE LAWS,UNC...|   23|
       M|             UNKNOWN|PUBLIC ADMINISTAT...|   22|
       M|      WHITE HISPANIC|LEAVING THE SCENE...|   22|
       M|AMERICAN INDIAN/A...|CRIMINAL MISCHIEF...|    4|
       U|             UNKNOWN|CRIMINAL MISCHIEF...|   11|
       M|               BLACK|    RESISTING ARREST| 1017|
       F|               BLACK|THEFT OF SERVICES...|  246|
       F|               BLACK|NY STATE LAWS,UNC...|   54|
       M|      WHITE HISPANIC| OBSTR BREATH/CIRCUL|  683|
       M|ASIAN / PACIFIC I...|MENACING,UNCLASSI...|  317|
       M|ASIAN / PACIFIC I...|LEAVING THE SCENE...|    8|
       M|ASIAN / PACIFIC I...|PUBLIC SAFETY,UNC...|    5|
       U|AMERICAN INDIAN/A...|ROBBERY,OPEN AREA...|    3|
       M|AMERICAN INDIAN/A...|CRIMINAL MISCHIEF...|   12|
       F|ASIAN / PACIFIC I...|ROBBERY,OPEN AREA...|   19|
       F|      WHITE HISPANIC|BURGLARY,UNCLASSI...|  131|
       M|      WHITE HISPANIC|RECKLESS ENDANGER...|  383|
       M|      WHITE HISPANIC|THEFT OF SERVICES...|   10|
```

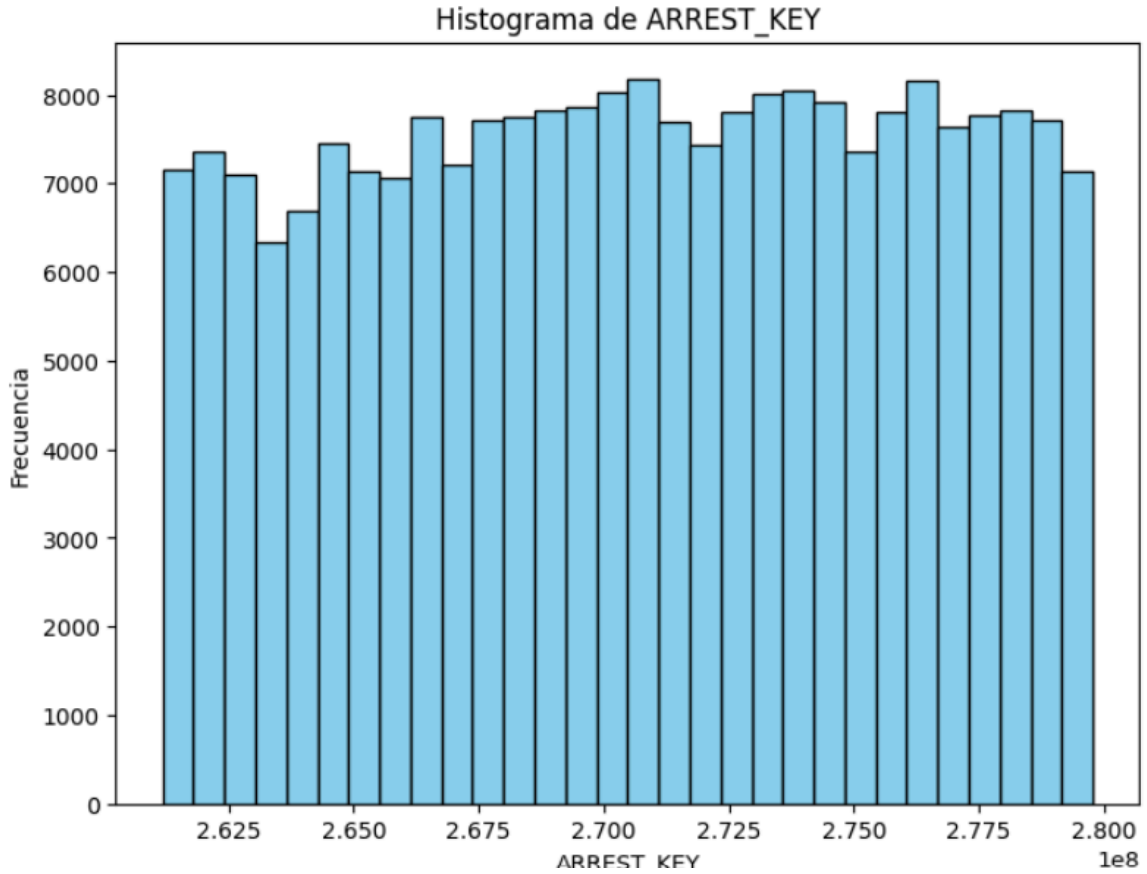
Interpretación:

- La mayoría de los perpetradores (50%) tienen entre 20 y 25 años.
- Hay una cantidad considerable de perpetradores (30%) menores de 20 años.
- Hay una cantidad menor de perpetradores (20%) mayores de 25 años.
- La asimetría a la derecha indica que hay una mayor concentración de perpetradores jóvenes.
- La dispersión amplia sugiere que existe una gran variabilidad en la edad de los

perpetradores.

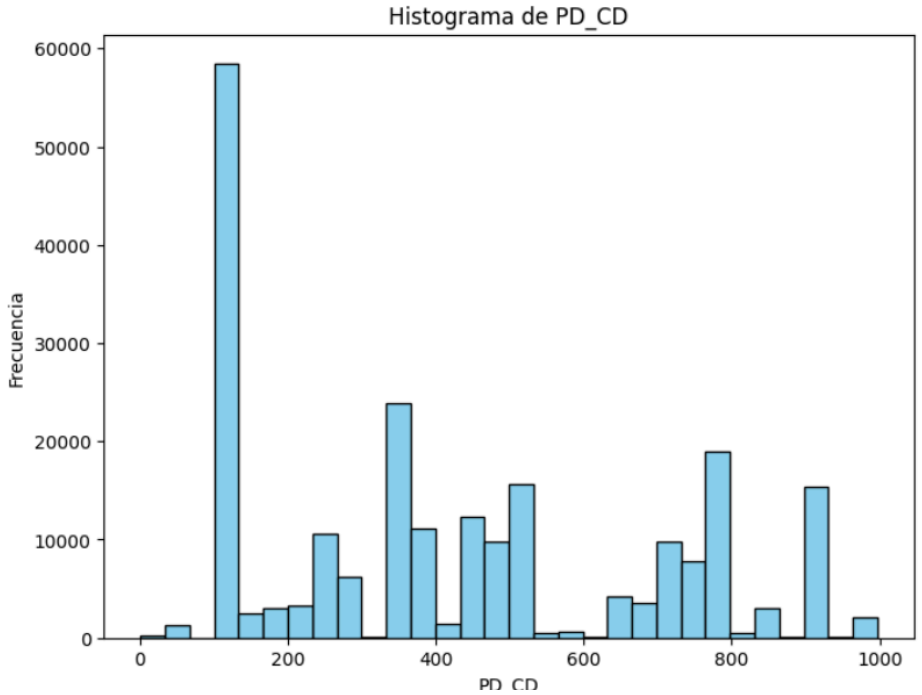
Se grafica un histograma para la columna `ARREST_KEY` y de esta manera visualizar la distribución de los datos con su frecuencia.

Resultado:



observamos los rangos de los valores que puede tomar ARREST_KEY

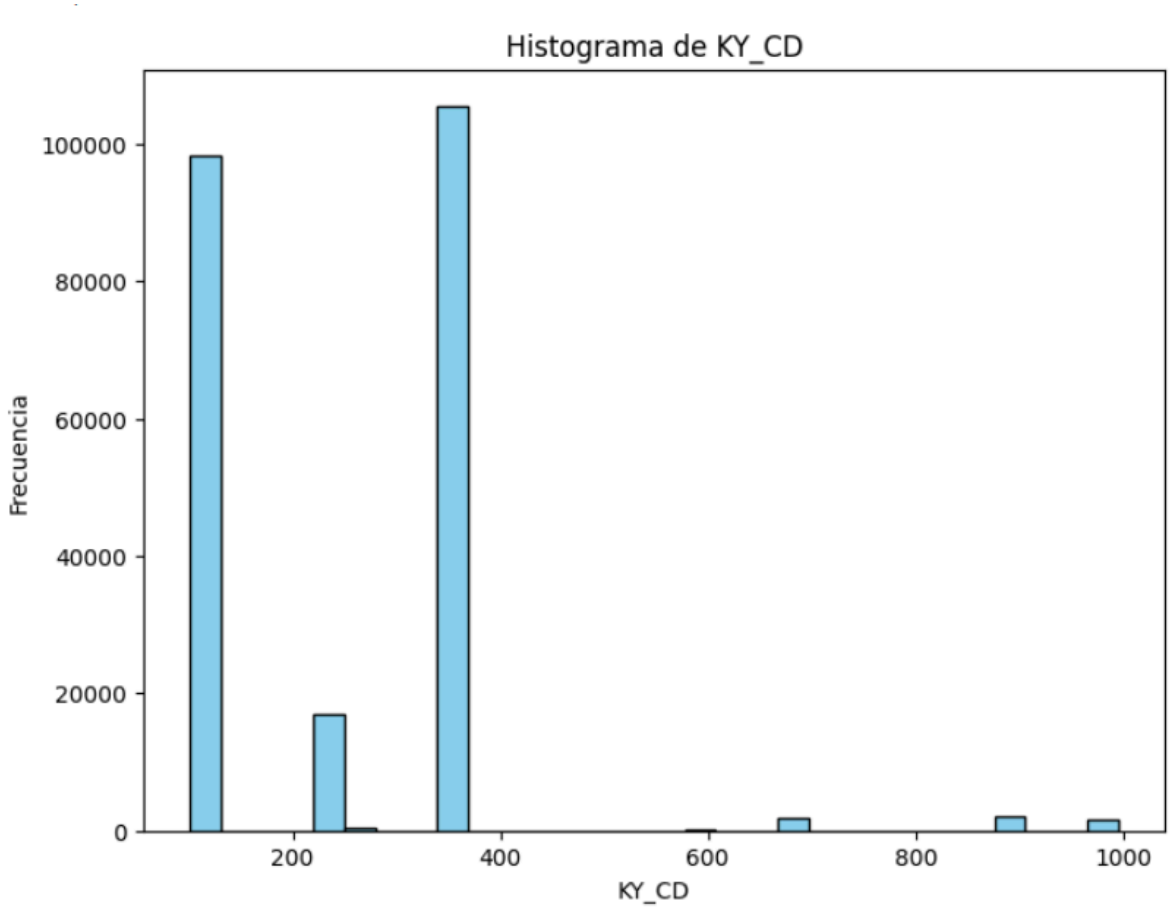
Variable `PD_CD`



Interpretación de la gráfica:

- La mayoría de las observaciones (50%) se encuentran entre 200 y 400 unidades de PD_CD.
- Hay una cantidad considerable de observaciones (30%) por debajo de 200 unidades de PD_CD.
- Hay una cantidad menor de observaciones (20%) por encima de 400 unidades de PD_CD.
- La asimetría a la derecha indica que hay una mayor concentración de valores en la parte inferior del rango de PD_CD.
- La dispersión amplia sugiere que existe una gran variabilidad en los valores de PD_CD.

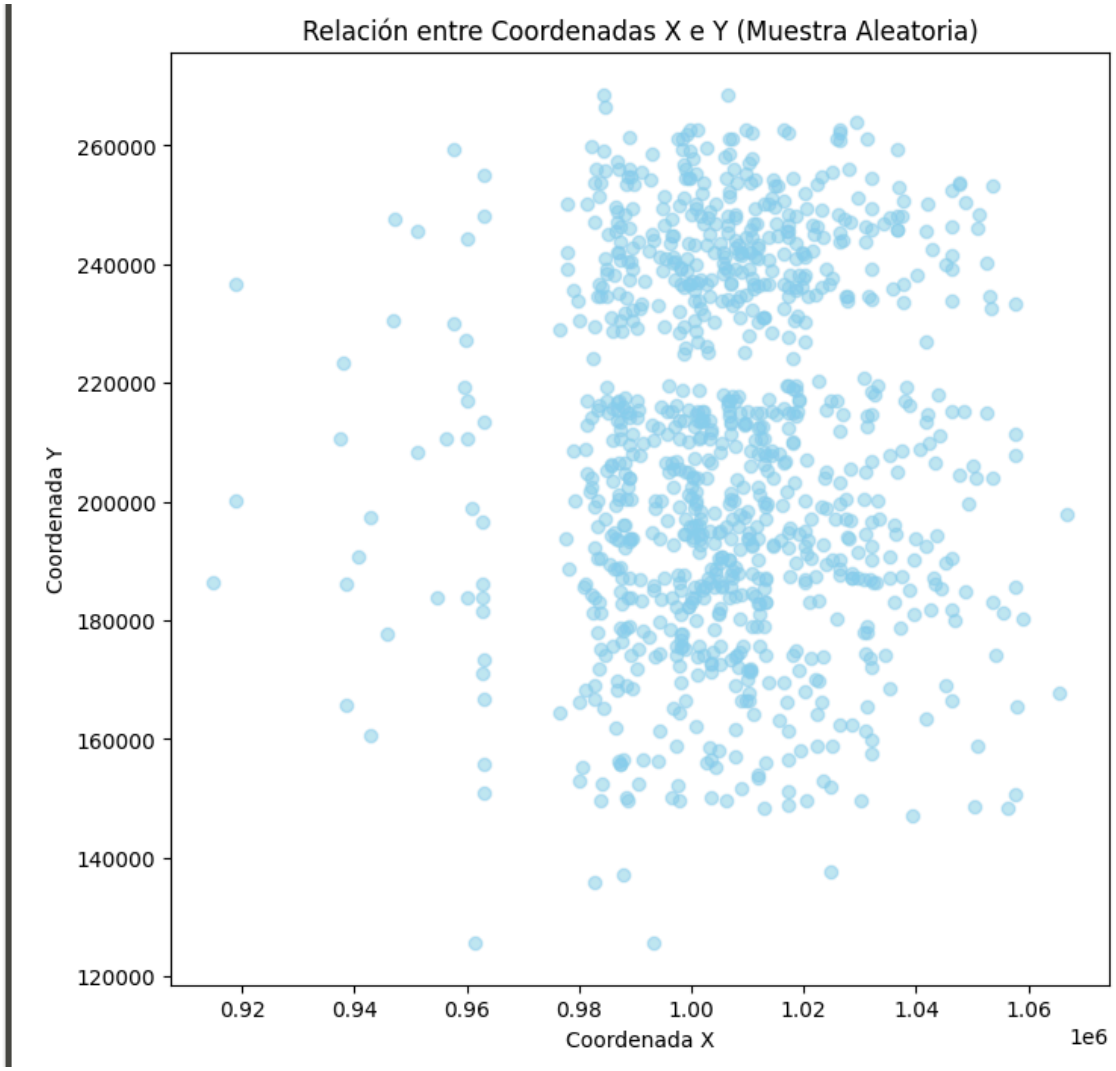
Variable KY_CD



Interpretación:

- El tipo de delito más común es la posesión de armas, etc.
- Los delitos de daño criminal y vandalismo también son relativamente comunes.
- Los demás tipos de delitos son menos comunes.
- La asimetría a la derecha indica que hay una mayor concentración de delitos en la categoría "POSSESSION OF WEAPONS, ETC.".
- La dispersión amplia sugiere que existe una gran variabilidad en los tipos de delitos que se están cometiendo.

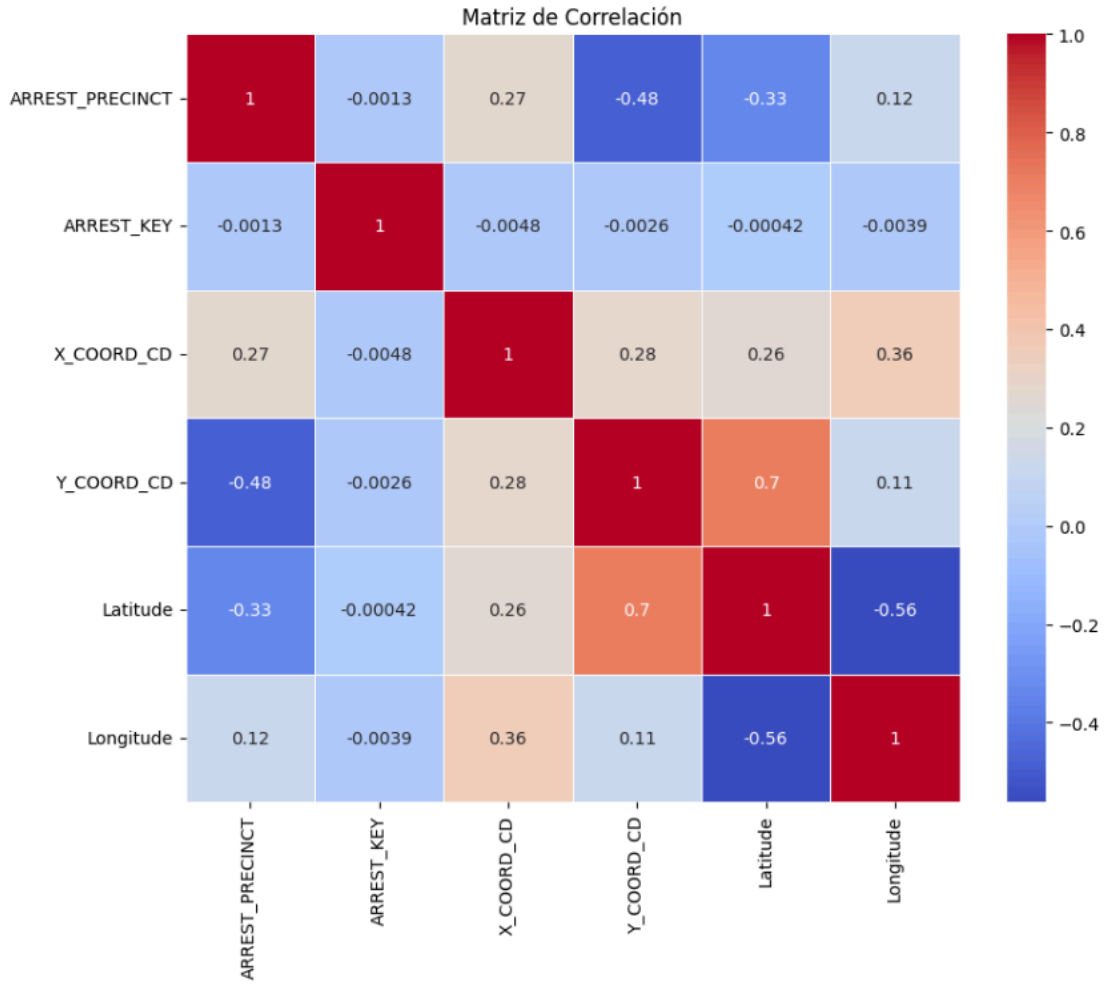
Variables: Relación entre Coordenadas X e Y (Muestra Aleatoria)



Interpretación:

- Fuerte correlación entre las coordenadas X y Y
- Pocos valores atípicos.
- Mayor concentración de los puntos en unas coordenadas específicas lo que nos puede decir que existen zonas de accidentes frecuente**s.**

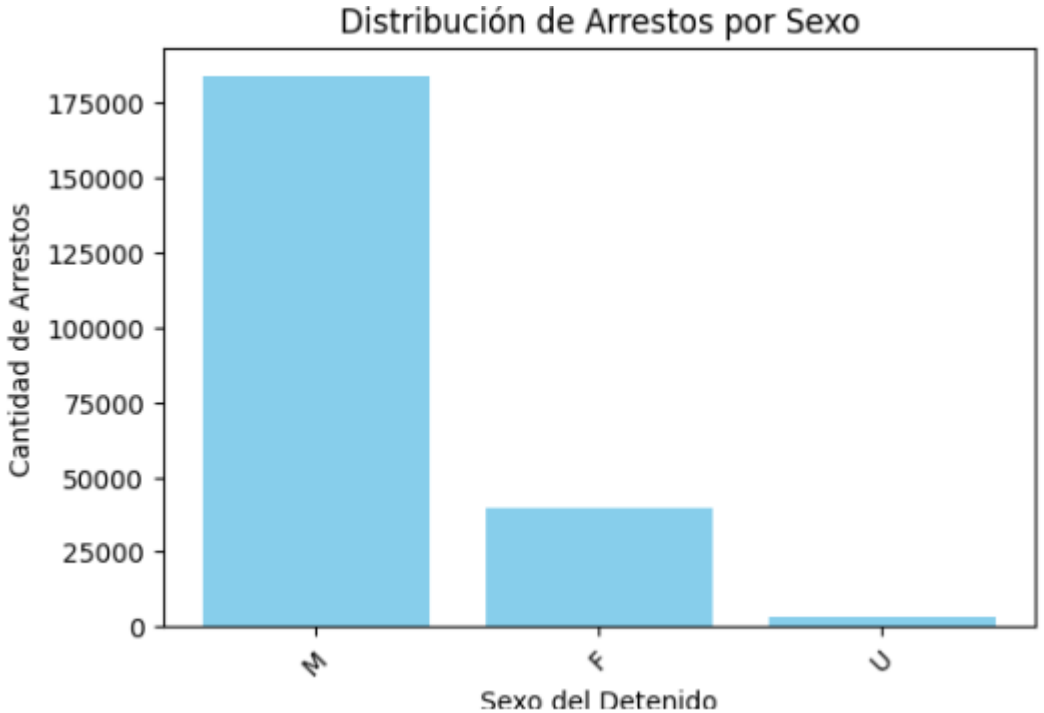
Matriz de correlación



Interpretación:

- ARREST_PRECINCT: No hay correlaciones fuertes con otras variables.
- ARREST_KEY: Correlaciones positivas débiles con X_COORD_CD, Y_COORD_CD y Longitude.
- X_COORD_CD: Correlación positiva moderada con Y_COORD_CD. Correlaciones positivas débiles con Latitude y Longitude.
- Y_COORD_CD: Correlación positiva moderada con X_COORD_CD. Correlación negativa débil con Latitude.
- Latitude: Correlación negativa débil con Y_COORD_CD. Correlación positiva débil con Longitude.
- Longitude: Correlaciones positivas débiles con X_COORD_CD, Y_COORD_CD y Latitude.

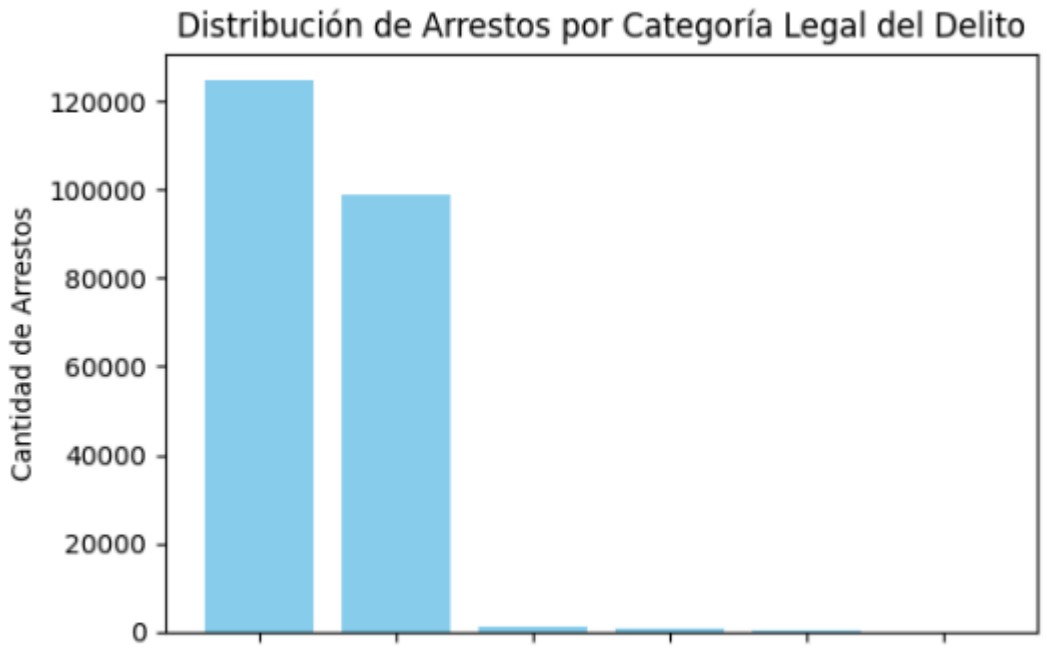
Distribución de Arrestos por Sexo



Interpretación General:

El gráfico de barras presentado muestra que la cantidad de arrestos es significativamente mayor para los hombres que para las mujeres. Sin embargo, la falta de información sobre el contexto de los arrestos limita la interpretación precisa del gráfico. Se necesita más información para comprender los factores que pueden estar contribuyendo a la diferencia en la cantidad de arrestos entre hombres y mujeres.

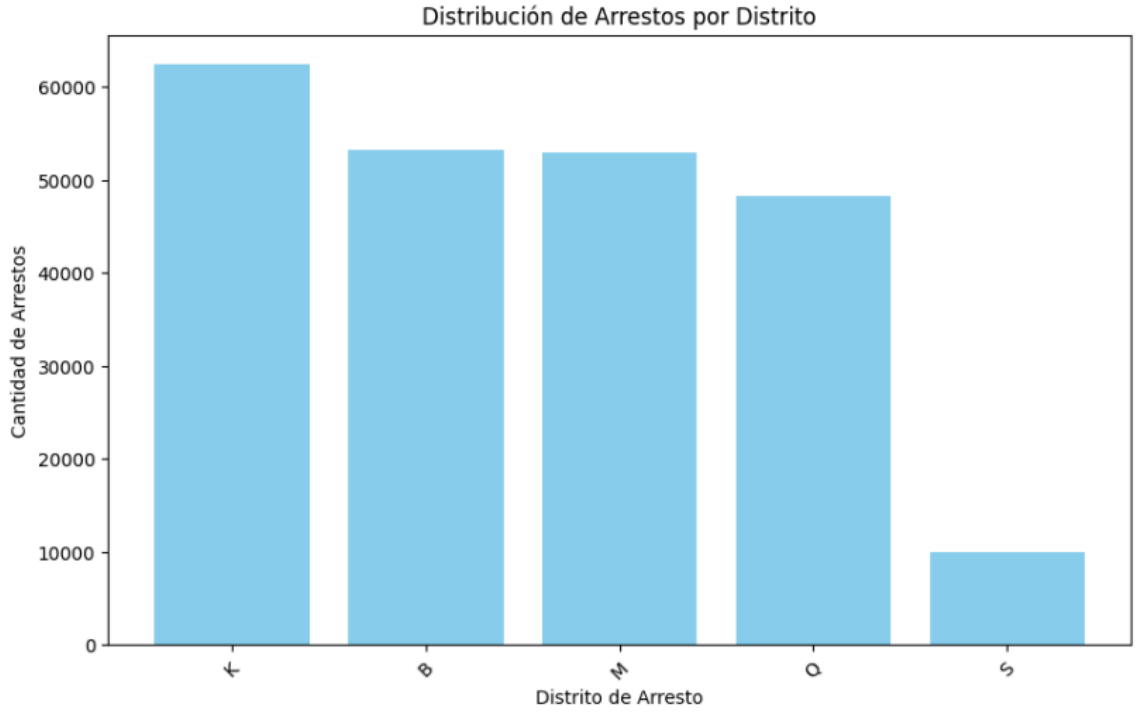
Distribución de Arrestos por Sexo



Interpretación General:

El gráfico de barras presentado muestra que la cantidad de arrestos varía considerablemente entre las diferentes categorías legales del delito. Las categorías con mayor cantidad de arrestos son "Delitos contra la Propiedad" y "Delitos contra la Seguridad Pública", mientras que las categorías con menor cantidad de arrestos son "Delitos contra la Libertad Sexual" y "Delitos contra la Administración Pública".

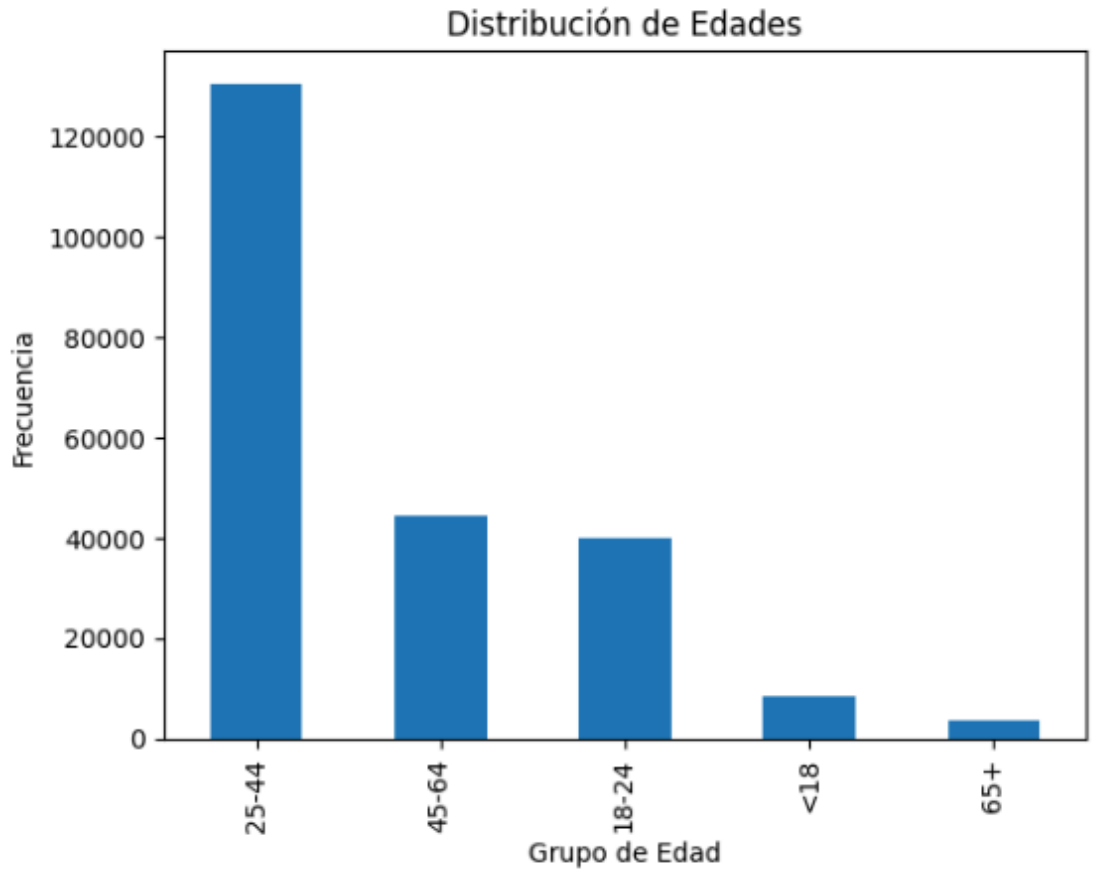
Distribución de Arrestos por Distrito



Interpretación:

- Cantidad de Arrestos: La cantidad de arrestos varía considerablemente entre los diferentes distritos.
- Distritos con mayor cantidad de arrestos: El distrito con mayor cantidad de arrestos es Brooklyn (K), seguido por Bronx (B), Queens (Q), Manhattan (M) y Staten Island (S).
- Distritos con menor cantidad de arrestos: El distrito con menor cantidad de arrestos es Staten Island (S).

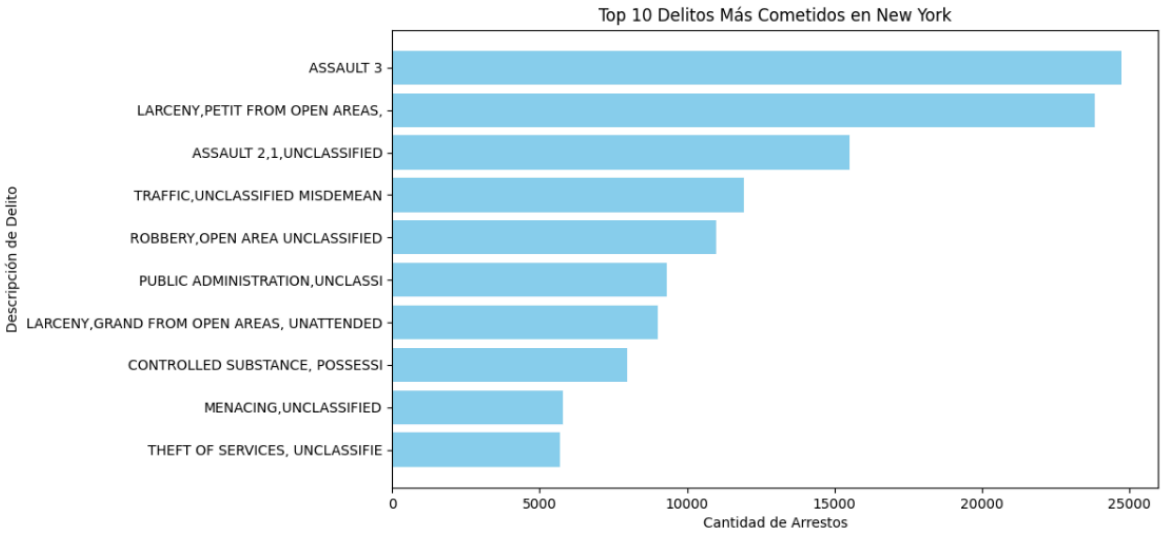
Arrestos por Distribución de Edades



Interpretación:

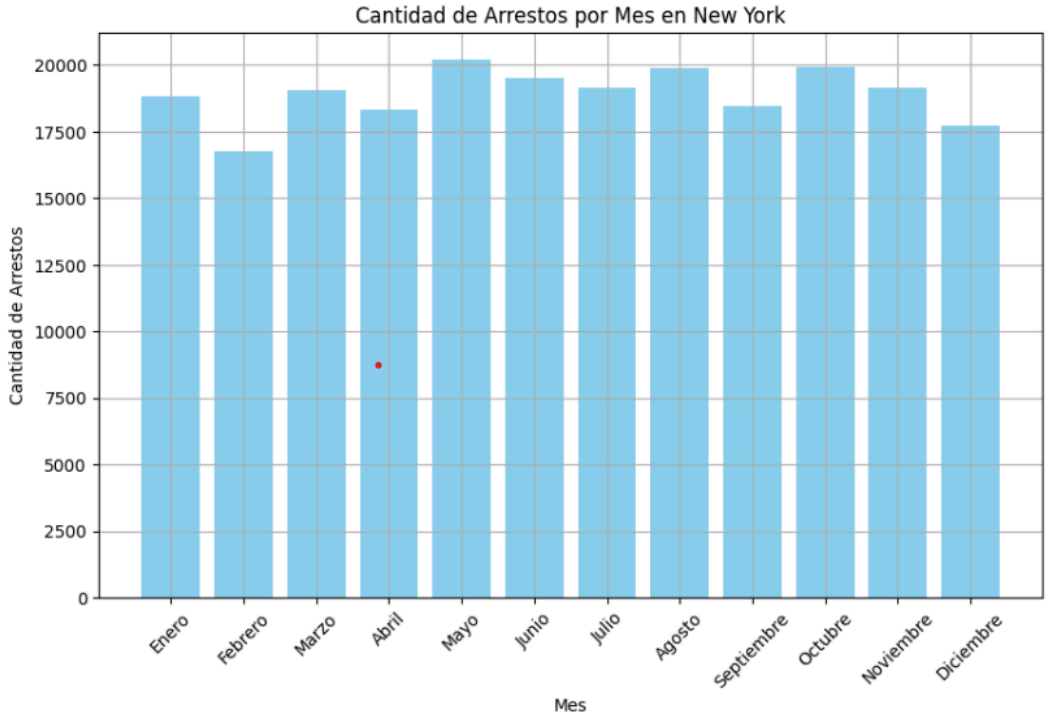
- Distribución de la Población: La población se distribuye de manera desigual entre los diferentes grupos de edad.
- Grupos de Edad con mayor Población: El grupo de edad con mayor población es el de 25 a 44 años.
- Grupos de Edad con menor Población: Los grupos de edad con menor población son los de menores de 18 años y mayores de 65 años.

Top 10 Delitos Más Cometidos en New York



Interpretación:
El delito más cometido, representado por la barra más larga, es "ASSAULT 3" con alrededor de 25,000 arrestos. Los otros delitos listados en el gráfico van disminuyendo en cantidad de arrestos, con el décimo delito más cometido teniendo aproximadamente entre 5,000 y 6,000 arrestos.

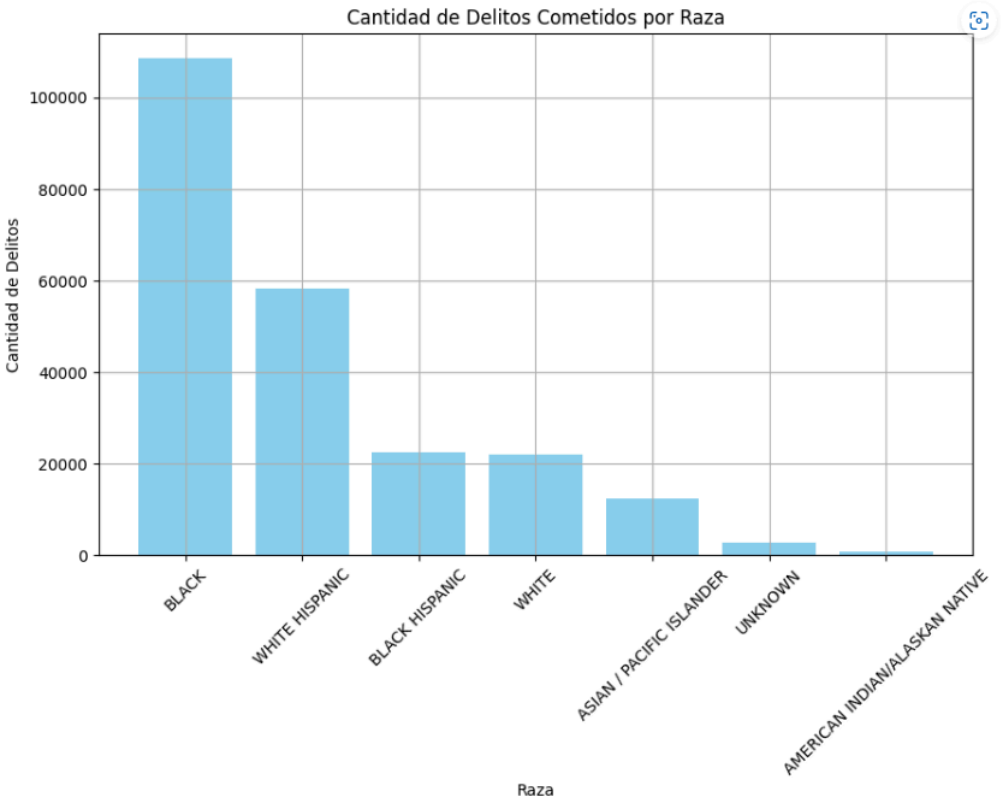
Cantidad de arrestos por mes



Interpretación:

- Hay una considerable variación en la cantidad de arrestos a lo largo de los diferentes meses. Los meses con mayor cantidad de arrestos parecen ser mayo, junio y julio.
- Los meses con menor cantidad de arrestos son noviembre, diciembre y febrero.
- Hay un pico muy pronunciado en el mes de mayo, que parece ser el mes con la mayor cantidad de arrestos registrados.

Cantidad de delitos cometidos por raza

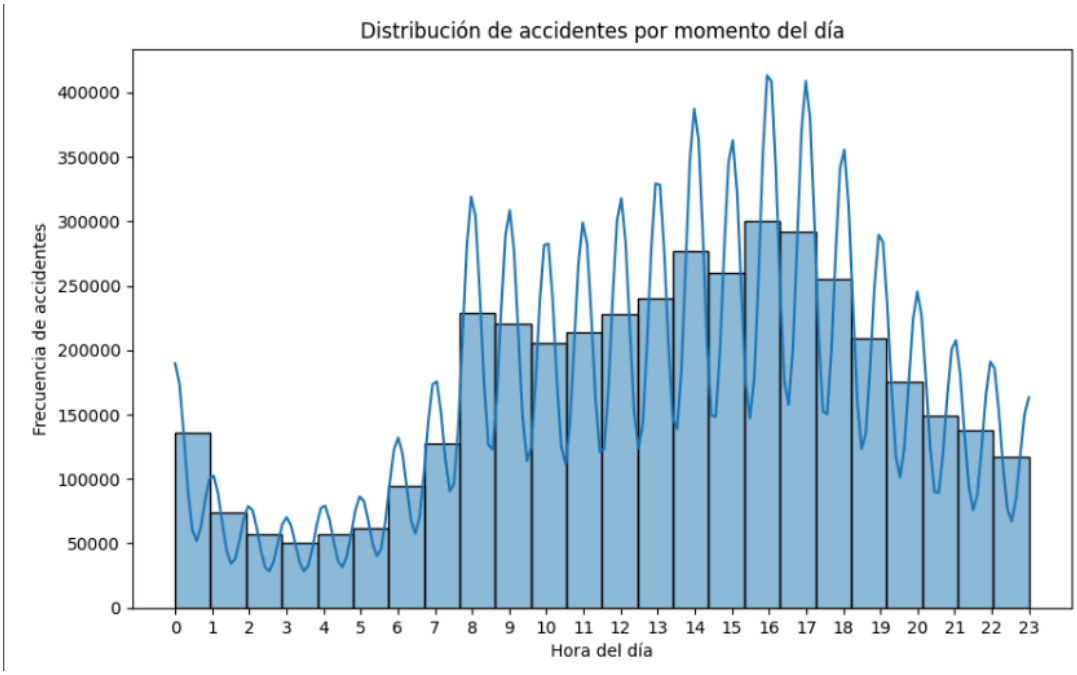


Interpretación:

- La raza con el mayor número de delitos cometidos es la raza negra (BLACK), con una cantidad muy superior a las demás razas.
- La segunda raza con más delitos cometidos es la raza hispana/latina (WHITE HISPANIC), seguida de la raza blanca (WHITE).
- Las razas con menor cantidad de delitos cometidos son las categorías "ASIAN / PACIFIC ISLANDER", "UNKNOWN" y "AMERICAN INDIAN/ALASKAN NATIVE".
- La gráfica muestra una gran disparidad en la cantidad de delitos cometidos entre las diferentes razas representadas.

Colisiones de vehículos motorizados - Vehículos

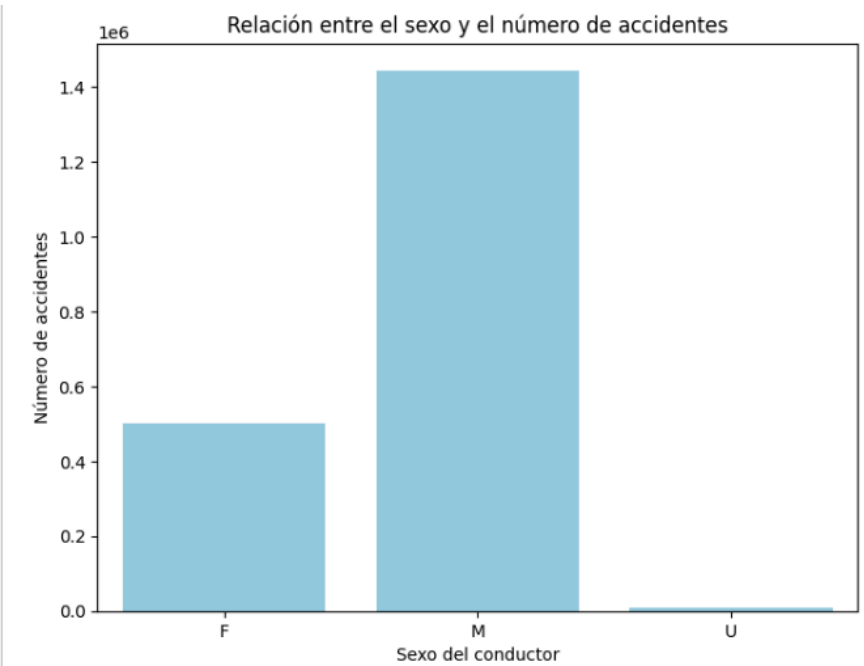
Distribución de accidente por hora del dia



Interpretación:

- Hay un patrón claramente definido a lo largo del día, con picos y valles que se repiten diariamente.
- Los momentos con mayor frecuencia de accidentes se observan durante las horas pico de la mañana, alrededor de las 8-9 am, y en las horas pico de la tarde, entre las 5-7 pm.
- Hay un periodo de baja actividad durante la madrugada, entre la 1 am y las 5 am aproximadamente, cuando se registra la menor cantidad de accidentes.
- Existen varios picos a lo largo del día, con varios momentos en los que la frecuencia de accidentes aumenta considerablemente, probablemente relacionados con patrones de tráfico y actividad humana.

Relacion entre el sexo y numero de accidentes



Interpretación:

- La categoría "M" (masculino) tiene un valor mucho más alto que las otras dos categorías, indicando que los conductores de sexo masculino tienen un mayor número de accidentes.
- La categoría "F" (femenino) tiene un valor significativamente más bajo que la categoría "M", sugiriendo que los conductores de sexo femenino tienen un menor número de accidentes.
- La categoría "U" (desconocido) tiene un valor muy bajo, casi insignificante en comparación con las otras dos.

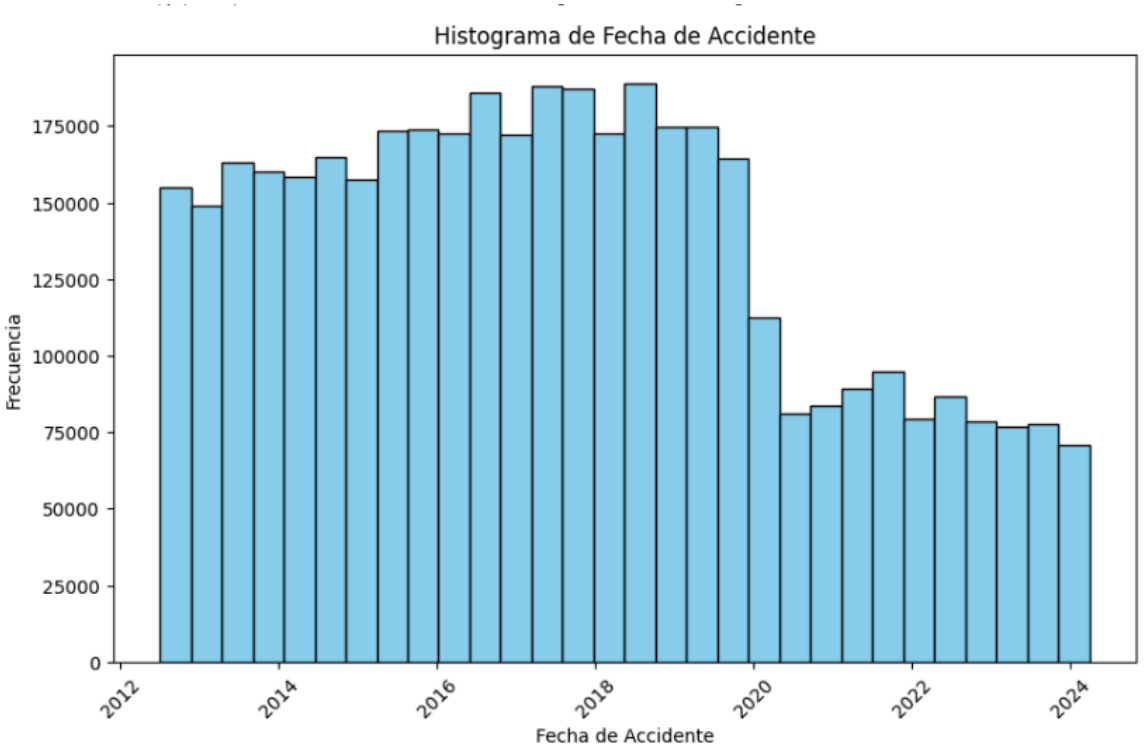
Número de accidentes por ocupantes en el vehículo

Ocupantes_Categoria	count
4 Ocupantes	37934
Más de 5 Ocupantes	12349
3 Ocupantes	91083
2 Ocupantes	317069
Desconocido	2190359
5 Ocupantes	13872
1 Ocupante	1506176

Interpretación:

- La categoría con mayor número de viviendas es la de "Desconocido" con 2,190,359 resultados.
- La siguiente categoría más numerosa es la de "1 Ocupantes" con 1,506,176 resultados.
- La tercera categoría más numerosa es la de "2 Ocupantes" con 317,069 resultados.
- La categoría "3 Ocupante" con 91,083 resultados.
- La categoría "4 Ocupantes" cuenta con 37,934 viviendas, mientras que la de "Más de 5 Ocupantes" tiene 13, 872 resultados.

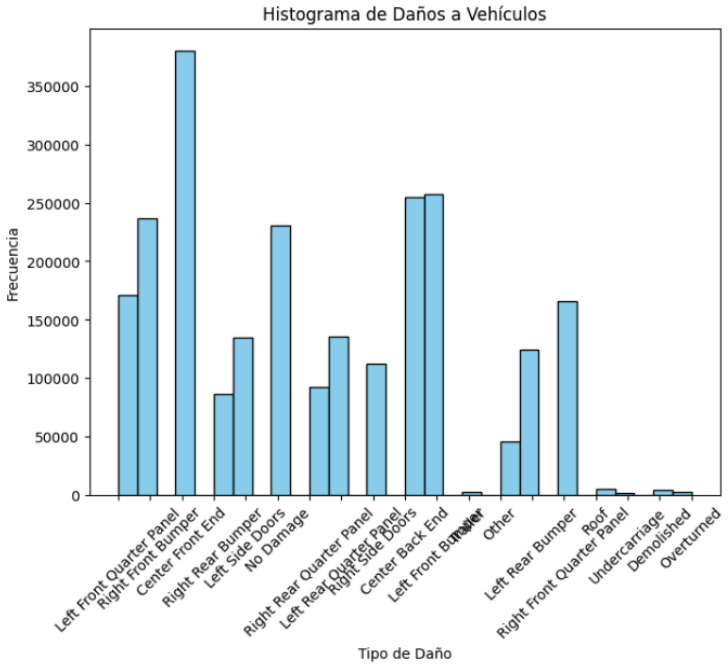
Cantidad de accidentes por año



Interpretación:

- Los años 2016, 2018 y 2019 parecen tener las frecuencias de accidentes más altas, mientras que 2020, 2022 y 2024 muestran una disminución significativa.
- La tendencia general a lo largo de los años parece ser un descenso gradual en la frecuencia de accidentes, aunque con fluctuaciones año a año.
- Es importante analizar los factores que pueden estar influyendo en estos patrones, como cambios en las regulaciones, mejoras en la seguridad, o variaciones en el volumen de tráfico y actividad económica que puedan afectar la ocurrencia de accidentes.

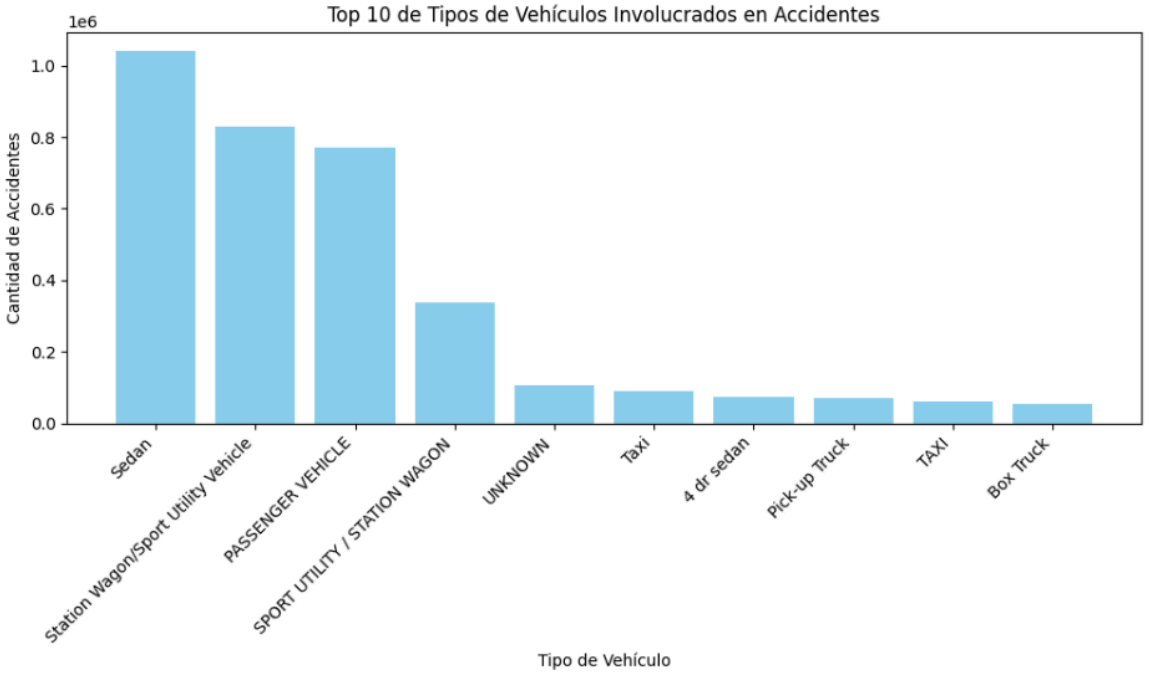
Histograma de Daños a Vehículos



Interpretación:

- El tipo de daño con mayor frecuencia es "Left Front Quarter Panel", lo que sugiere que este es uno de los tipos de daño más común.
- Otros tipos de daño significativos incluyen "Left Rear Quarter Panel", "Right Front Quarter Panel" y "Right Rear Quarter Panel", lo que indica que los daños en las partes delanteras y traseras de los vehículos son bastante frecuentes.
- Hay algunos tipos de daño menos comunes, como "Undercarriage", "Derailed" y "Other", que representan una fracción mucho menor de los incidentes.
- La distribución general muestra una variación considerable en la frecuencia de los diferentes tipos de daños, lo que puede ser útil para comprender los patrones de siniestros y enfocar los esfuerzos de prevención.

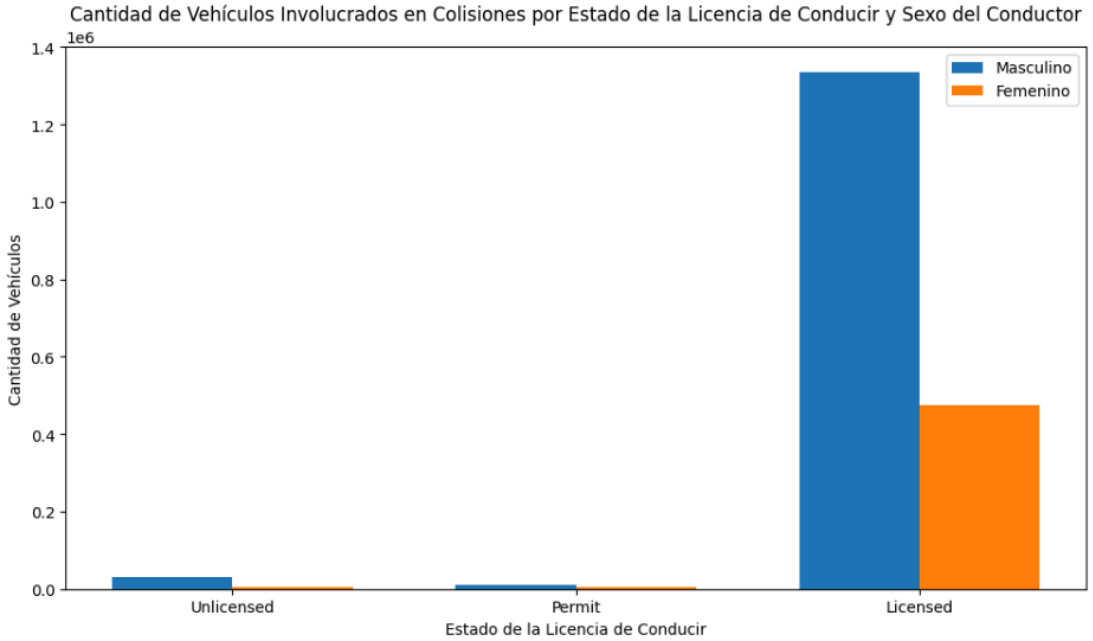
Top 10 vehículos involucrados en accidentes



Interpretación:

- El tipo de vehículo con mayor participación en accidentes es el Sedan, con una frecuencia notablemente más alta que el resto.
- Otros tipos de vehículos con alta participación incluyen los Sport Utility Vehicle (SUV), los Passenger Vehicle, y los Sport Utility/Station Wagon.
- Algunos tipos de vehículos menos frecuentes en los accidentes son las Pick-up Truck, las 4 dr sedan, las Taxi y los Box Truck.
- La gráfica permite identificar los tipos de vehículos que requieren una mayor atención en términos de seguridad y prevención de accidentes, ya que su alta frecuencia puede implicar riesgos más elevados.
- Esta información puede ser valiosa para enfocar esfuerzos de mejora en diseño, tecnología, capacitación de conductores, entre otros, en los tipos de vehículos más propensos a verse involucrados en siniestros.

Estado de licencia por sexo



Interpretación;

- La mayoría de los vehículos involucrados en colisiones son conducidos por personas con licencia.
- Los hombres tienen una mayor participación en colisiones que las mujeres, en todas las categorías de licencia.
- La categoría de sin licencia tiene la menor cantidad de vehículos involucrados en colisiones, tanto para hombres como para mujeres.

4. **Reporte calidad de datos:**

Reporte de calidad de datos sobre arrestos:

```python
from pyspark.sql.functions import count, when, isnan

# Luego puedes usar count y otras funciones sin problemas
df1.select([count(when(col(c).isNull(), c)).alias(c) for c in df1.columns]).show()
```

ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW_CODE	LAW_CAT_CD	ARREST_BORO	ARREST_PRECINCT	JURISDICTION_CODE
0	0	2	0	17	0	0	1599	0	0	0

AGE_GROUP	PERP_SEX	PERP_RACE	X_COORD_CD	Y_COORD_CD	Latitude	Longitude	New Georeferenced Column	ARREST_YEAR	ARREST_MONTH
0	0	0	0	0	0	0	0	0	0

Análisis de Valores Faltantes en Datos de Arrestos

En este apartado, se muestra el conteo de valores faltantes para la tabla de datos de arrestos. Se identifican varias columnas con valores faltantes, como PD_CD (2 nulos), KY_CD (17 nulos) y LAW_CAT_CD (1599 nulos). Estos hallazgos son cruciales ya que los valores faltantes pueden impactar la calidad y validez de nuestro análisis.

Estrategias para Manejar Valores Faltantes:

- PD_CD y KY_CD: Estas columnas representan variables categóricas con códigos. Una opción sería reemplazar los valores nulos por una nueva categoría que indique "Código desconocido". Esto nos permite conservar la información sobre la ausencia de datos sin perder la integridad del conjunto.
- LAW_CAT_CD: Esta variable categórica también presenta valores faltantes. Aquí, podríamos aplicar la técnica de imputación basada en la moda. Esto implica reemplazar los valores faltantes por la categoría legal más frecuente observada en los registros no nulos. De esta manera, mantenemos la coherencia en nuestros datos y minimizamos la pérdida de información.

Reporte de calidad de datos sobre accidentes vehiculares:

```python
df3.select([count(when(col(c).isNull(), c)).alias(c) for c in df3.columns]).show()
```

UNIQUE_ID	COLLISION_ID	CRASH_DATE	CRASH_TIME	VEHICLE_ID	STATE_REGISTRATION	VEHICLE_TYPE	VEHICLE_MAKE	VEHICLE_MODEL	VEHICLE_YEAR	TRAVEL_DIRECTION	VEHICLE_OCCUPANTS	DRIVER_SEX
0	0	0	0	0	302396	235232	1878192	4117412	1897871	1666966	1780789	2215737

DRIVER_LICENSE_STATUS	DRIVER_LICENSE_JURISDICTION	PRE_CRASH	POINT_OF_IMPACT	VEHICLE_DAMAGE	VEHICLE_DAMAGE_1	VEHICLE_DAMAGE_2	VEHICLE_DAMAGE_3	PUBLIC_PROPERTY_DAMAGE	PUBLIC_PROPERTY_DAMAGE_TYPE	CONTRIBUTING_FACTOR_1	CONTRIBUTING_FACTOR_2	CRASH_HOUR
2304191	2299404	920140	1699970	1724222	2594754	2983615	3260601	1528858	4142850	147285	1686788	0

Al examinar la tabla de conteos de valores nulos, se identifican varias columnas con una cantidad considerable de registros faltantes. Por ejemplo, las columnas STATE_REGISTRATION, VEHICLE_MAKE, VEHICLE_MODEL, VEHICLE_YEAR, TRAVEL_DIRECTION, VEHICLE_OCCUPANTS, DRIVER_SEX,

DRIVER_LICENSE_STATUS, DRIVER_LICENSE_JURISDICTION, PRE_CRASH, POINT_OF_IMPACT, VEHICLE_DAMAGE, VEHICLE_DAMAGE_1, VEHICLE_DAMAGE_2, VEHICLE_DAMAGE_3, PUBLIC_PROPERTY_DAMAGE, PUBLIC_PROPERTY_DAMAGE_TYPE, CONTRIBUTING_FACTOR_1 y CONTRIBUTING_FACTOR_2 presentan un número significativo de valores faltantes.

Estas ausencias de datos podrían deberse a diversos factores, como errores en la recopilación de información en el lugar del accidente, falta de cumplimentación de ciertos campos en los formularios, problemas en la integración de datos de diferentes fuentes, o simplemente información no disponible para determinados registros.

Para tratar estos valores faltantes, se podrían aplicar diferentes técnicas dependiendo de la naturaleza de los datos y el contexto específico. Por ejemplo, para variables categóricas como VEHICLE_TYPE, DRIVER_SEX o DRIVER_LICENSE_STATUS, se podría asignar una categoría especial para los valores faltantes o utilizar técnicas de imputación basadas en la moda. Para variables numéricas como VEHICLE_OCCUPANTS o VEHICLE_YEAR, se podrían utilizar métodos de imputación como la media, la mediana o modelos de regresión. Otra opción sería eliminar los registros con valores faltantes, aunque esto podría llevar a una pérdida significativa de información si la cantidad de registros afectados es alta.

Planteamiento de preguntas

a. ¿Es frecuente que los autos tengan daños en lugares específicos (VEHICLE_DAMAGE, VEHICLE_DAMAGE_1, VEHICLE_DAMAGE_2, VEHICLE_DAMAGE_3) después de un accidente?

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

c. ¿Los accidentes tienden a ocurrir más en alguna dirección de viaje específica (TRAVEL_DIRECTION)?

d. ¿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?

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

f. ¿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*)?

g. ¿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))?

h. ¿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?

i. ¿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?

Filtros, limpieza y transformación inicial:

1. **Datos de arrestos del Departamento de Policía de Nueva York hasta la fecha:**

En esta etapa, se realizaron diversas operaciones para garantizar la calidad y consistencia de los datos. Primero, se eliminaron los registros duplicados basados en las columnas clave ARREST_KEY, ARREST_DATE, PD_CD, PD_DESC, KY_CD, OFNS_DESC, LAW_CODE y LAW_CAT_CD, lo cual asegura que cada registro sea único y no se cuente más de una vez en los análisis posteriores.
Luego, se contabilizaron los valores nulos en cada columna del DataFrame utilizando la función count y when de PySpark. Esto permitió identificar las columnas con valores faltantes y tomar medidas apropiadas.
Para tratar los valores nulos, se aplicó una estrategia de imputación, reemplazando los valores faltantes en las columnas PD_CD y KY_CD con la etiqueta 'Desconocido'. Esta técnica evita la pérdida de registros completos y permite mantener la integridad de los datos.
Además, se identificó la necesidad de realizar un análisis de agrupamiento (clustering) para explorar la relación entre LAW_CODE (Código de la ley asociada al delito) y LAW_CAT_CD (Categoría legal del delito). Para facilitar este análisis, se planea llenar los valores nulos en la columna LAW_CAT_CD utilizando una estrategia de imputación basada en la información disponible en LAW_CODE.

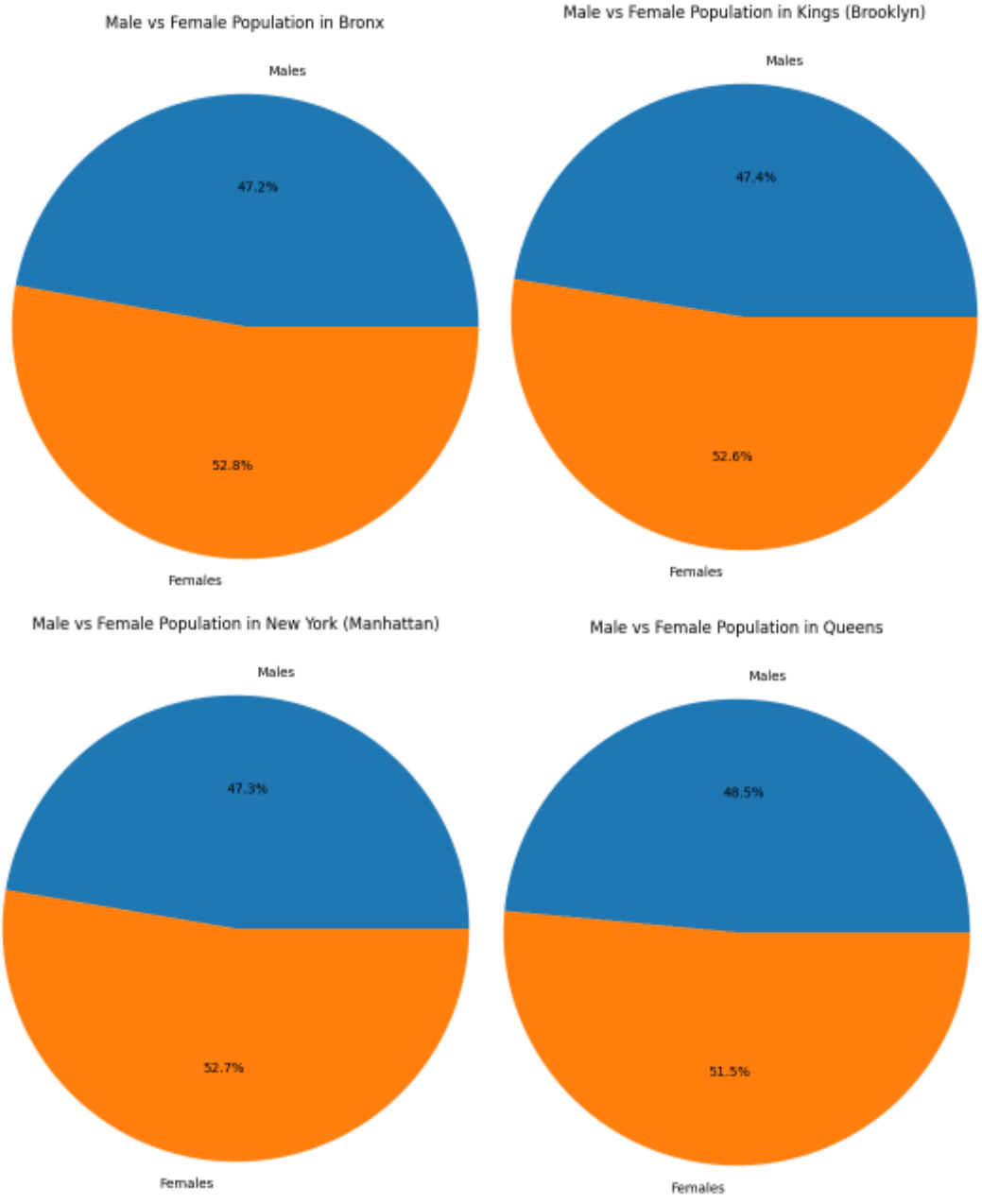
2. **Colisiones de vehículos motorizados - Vehículos:**

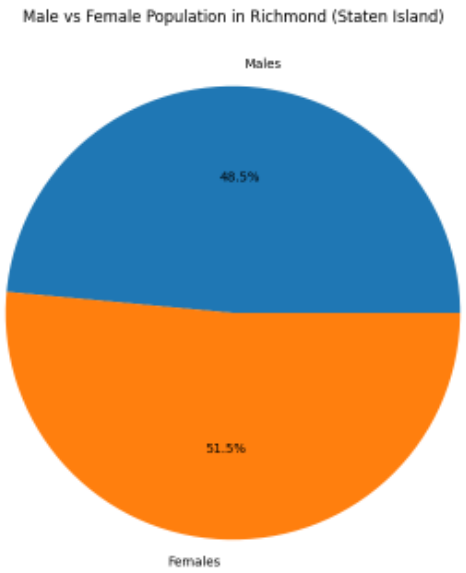
En primer lugar, se eliminaron los registros duplicados presentes en el DataFrame df3 utilizando el método dropDuplicates(). Esto asegura que cada registro sea único y evita contar información redundante en los análisis posteriores.
Luego, se contabilizaron los valores nulos en cada columna del DataFrame utilizando la función count y se realizaron diversas transformaciones para tratar estos valores faltantes.
Para la columna VEHICLE_OCCUPANTS, se convirtieron los valores numéricos a cadenas de texto y se creó una nueva columna Ocupantes_Categoria que categoriza el número de ocupantes de manera descriptiva (ej: "1 Ocupante", "2 Ocupantes", etc.). Los valores nulos se reemplazaron con "0" y se mapearon a la categoría "Desconocido".
Se imputaron valores 'Desconocido' en las columnas STATE_REGISTRATION, VEHICLE_MAKE, VEHICLE_MODEL, DRIVER_SEX y DRIVER_LICENSE_JURISDICTION para manejar los registros faltantes.
Además, se utilizó la moda (valor más frecuente) para imputar valores faltantes en VEHICLE_DAMAGE, PUBLIC_PROPERTY_DAMAGE y PUBLIC_PROPERTY_DAMAGE_TYPE.
Para las columnas PRE_CRASH, POINT_OF_IMPACT, CONTRIBUTING_FACTOR_1 y CONTRIBUTING_FACTOR_2, se planea buscar la probabilidad de cada valor y llenar los nulos en consecuencia.

Bono

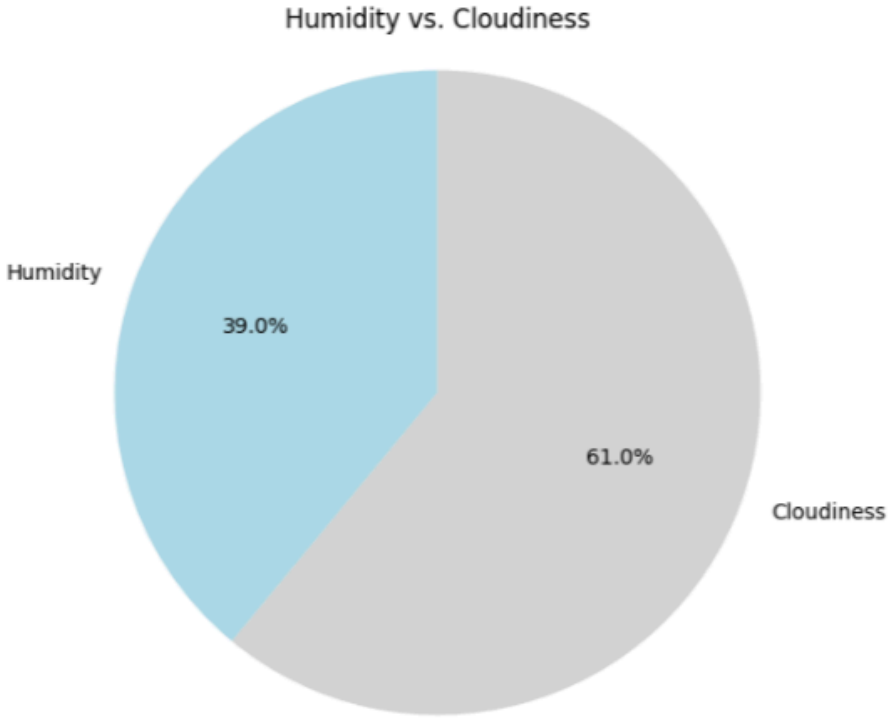
- **Web Scraping**

Para esta sección, empleamos la librería ´pandas´ de Python para extraer datos de una página web que contenía información valiosa en formato de tabla- La elección de ´pandas´´ y su método pd.read_html se debió a la estructura accesible y bien definida de los datos en la página web, lo que permitió una extracción directa y eficiente sin la necesidad de herramientas más complejas como Selenium. Esta técnica de Web Scraping fue crucial para obtener datos demográficos específicos, destacando un hallazgo interesante, el cual es que en general, el porcentaje de mujeres es mayor que el de los hombres en cada localidad examinada.





- **API**
  Para esta sección, se utilizó una API, con la cual pudimos acceder a datos climáticos actualizados y precisos, como la velocidad del viento, la humedad y la nubosidad. Esta información nos permitió llevar a cabo un análisis más detallado sobre las condiciones climáticas. También nos permitió extraer los datos pertinentes y así generar visualizaciones, entre ellos, generamos visualizaciones que exploran otras dimensiones climáticas como la velocidad del viento, la humedad y la nubosidad, en donde se puede ver que la nubosidad es mayor a la humedad(01-04-2024), sin embargo irá cambiando en el momento en que se llama.



Bibliografía:

Police Department (NYPD). (2024). NYPD Arrest Data (Year to Date). Recuperado de https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc/about_data

Police Department (NYPD). (2024). Motor Vehicle Collisions - Vehicles. Recuperado de https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4/about_data

Ciudad de Nueva York. (2024). Vision Zero NYC. [Página web]. https://www.nyc.gov/visionzero

National Geographic Viajes. (s.f.). National Geographic Viajes [Página web de viajes y destinos turísticos]. https://viajes.nationalgeographic.com.es/c/nueva-york