

Procesamiento de Datos A Gran Escala: Colisiones & Arrestos

Por Daniela Torres, Daniel Sandoval e Isaac Janica

Índice

01

Entendimiento del negocio

02

Selección, colección y descripción de los datos

03

Análisis Exploratorio

04

Reporte de calidad

05

Planteamiento de preguntas

06

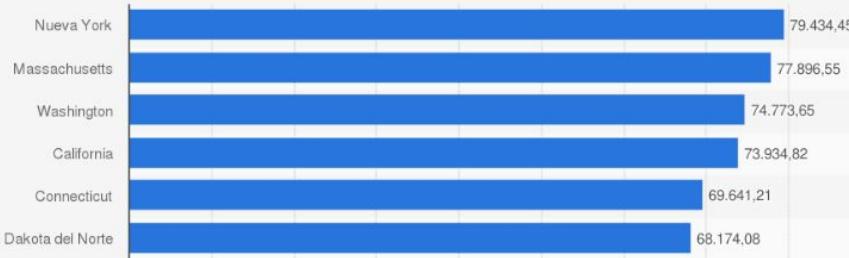
Filtro limpieza y transformación

01

Entendimiento del Negocio

Datos Macroeconómicos y generales

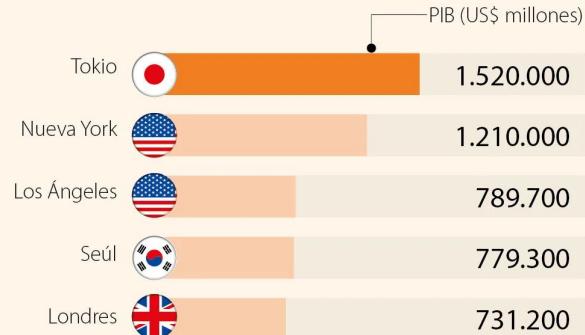
Producto Interior Bruto (PIB) real per cápita en los Estados Unidos en 2022, por Estado (en USD encadenados de 2012)



NYC cómo economía

New York es la ciudad con mayor PIB per cápita en los EE.UU y la segunda con el PIB más grande del mundo.

LAS 10 CIUDADES QUE POR SU PIB SON IGUALES A UN ESTADO

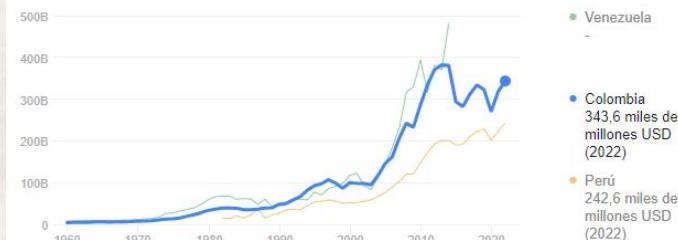


Fuente: un.org / wearetop10.com

Gráfico: LR-GR

Colombia / Producto interior bruto

343,6 miles de millones USD (2022)



NYC en términos de Seguridad

Index Crime Statistics: January 2024

	Jan. 2024	Jan. 2023	+/-	% Change
Murder	27	36	-9	-25.0%
Rape	102	135	-33	-24.4%
Robbery	1417	1345	72	5.4%
Felony Assault	2068	2100	-32	-1.5%
Burglary	1065	1328	-263	-19.8%
Grand Larceny	4056	4041	15	0.4%
Grand Larceny Auto	1178	1224	-46	-3.8%
TOTAL	9913	10209	-296	-2.9%

Hate Crimes Statistics: January 2024

(Representing January 1 – January 31 for calendar years 2024 and 2023)

Motivation	2024	2023	Diff	% Change
Asian	1	1	0	0%
Black	3	4	-1	-25%
Ethnic	2	1	1	100%
Gender	1	0	1	***
Hispanic	1	0	1	***
Jewish	31	17	14	82%
Muslim	0	1	-1	-100%
Religion	3	2	1	50%
Sexual Orientation	2	2	0	0%
White	1	5	-4	-80%
TOTAL	45	33	12	36%

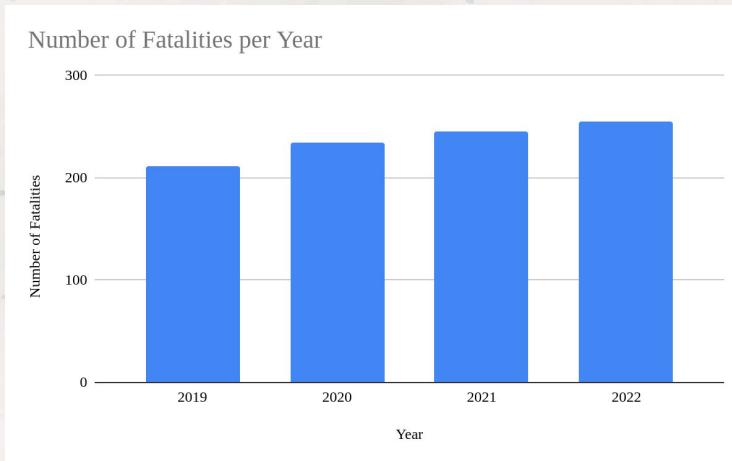
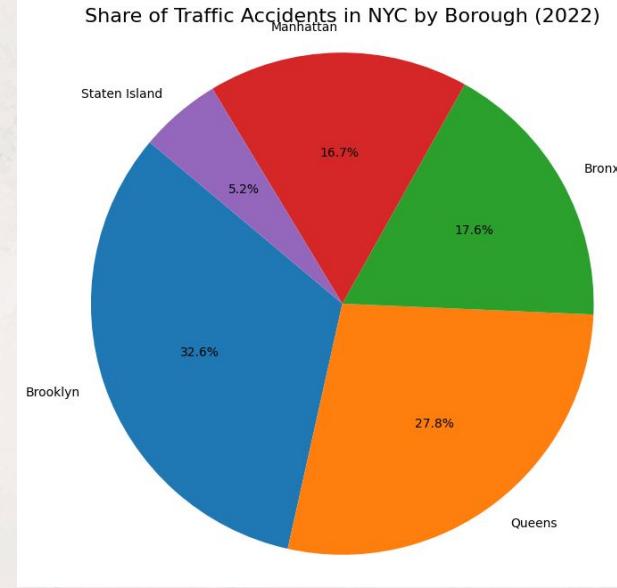
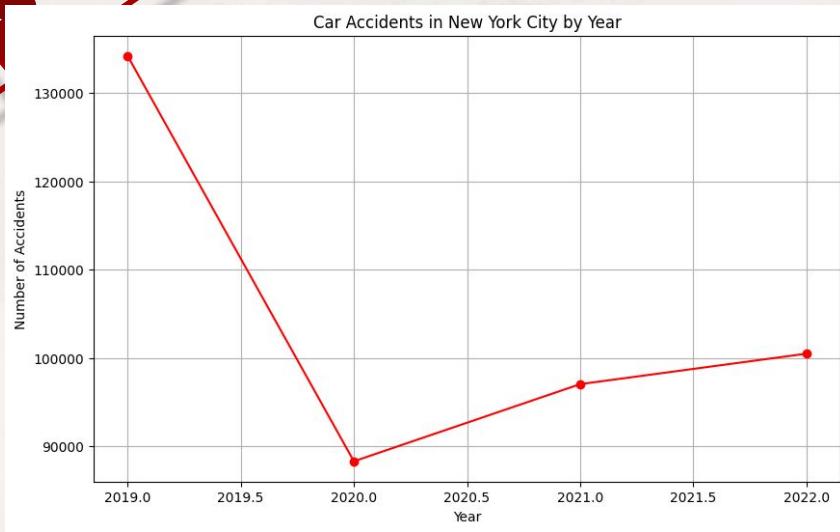
Hate Crimes Statistics: January 2024

(Representing January 1 – January 31 for calendar years 2024 and 2023)

Motivation	2024	2023	Diff	% Change
Asian	1	1	0	0%
Black	3	4	-1	-25%
Ethnic	2	1	1	100%
Gender	1	0	1	***
Hispanic	1	0	1	***
Jewish	31	17	14	82%
Muslim	0	1	-1	-100%
Religion	3	2	1	50%
Sexual Orientation	2	2	0	0%
White	1	5	-4	-80%
TOTAL	45	33	12	36%

	Jan. 2024	Jan. 2023	+/-	% Change
Murder	27	36	-9	-25.0%
Rape	102	135	-33	-24.4%
Robbery	1417	1345	72	5.4%
Felony Assault	2068	2100	-32	-1.5%
Burglary	1065	1328	-263	-19.8%
Grand Larceny	4056	4041	15	0.4%
Grand Larceny Auto	1178	1224	-46	-3.8%
TOTAL	9913	10209	-296	-2.9%

**NYC en
términos de
Seguridad**



**NYC en
términos de
Accidentes
Viales**

02

Selección, colección y descripción de los Datos

Qué utilizar, que tipos, su significado y descripción general

Colección y Descripción de los Datos

Nombre de la Columna	Descripción de la Columna	Tipo de Datos
ID_COLISIÓN	Código de registro único generado por el sistema	Cadena
FECHA_ACCIDENTE / FECHA_COLISIÓN	Fecha de ocurrencia de la colisión	Cadena
HORA_ACCIDENTE / HORA_COLISIÓN	Hora de ocurrencia de la colisión	Cadena
ID_ÚNICO	Código de registro único generado por el sistema	Cadena
ID_COLISIÓN	Código de identificación único del choque	Cadena
ID_VEHÍCULO	Código de identificación del vehículo asignado por el sistema	Cadena
TIPO_VEHÍCULO	Tipo de vehículo basado en la categoría seleccionada (ATV, bicicleta, automóvil/SUV, ebike, patinete eléctrico, camión/autobús, motocicleta, otro)	Cadena
MARCA_VEHÍCULO	Marca del vehículo	Cadena
MODELO_VEHÍCULO	Modelo del vehículo	Cadena
AÑO_VEHÍCULO	Año de fabricación del vehículo	Cadena
DIRECCIÓN_VIAJE	Dirección en la que se desplazaba el vehículo	Cadena
OCCUPANTES_VEHÍCULO	Número de ocupantes del vehículo	Cadena
SEXO_CONDUCTOR	Género del conductor	Cadena

Nombre de la Columna	Descripción de la Columna	Tipo de Datos
ESTADO_LICENCIA_CONDUTOR	Licencia, permiso, sin licencia	Cadena
JURISDICCIÓN_LICENCIA_CONDUTOR	Estado donde se emitió la licencia de conducir	Cadena
ACCIÓN_PRE_ACDNT	Ir recto, girar a la derecha, adelantar, retroceder, etc.	Cadena
PUNTO_DE_IMPACTO	Ubicación en el vehículo del punto de impacto inicial (es decir, lado del conductor, parte trasera del lado del pasajero, etc.)	Cadena
DAÑO_VEHÍCULO	Ubicación en el vehículo donde ocurrió la mayor parte del daño	Cadena
DAÑO_VEHÍCULO_1	Ubicaciones adicionales de daño en el vehículo	Cadena
DAÑO_VEHÍCULO_2	Ubicaciones adicionales de daño en el vehículo	Cadena
DAÑO_VEHÍCULO_3	Ubicaciones adicionales de daño en el vehículo	Cadena
DAÑO_PROPRIEDAD_PÚBLICA	Propiedad pública dañada (Sí o No)	Cadena
TIPO_DAÑO_PROPRIEDAD_PÚBLICA	Tipo de propiedad pública dañada (por ejemplo, señal, cerca, poste de luz, etc.)	Cadena
FACTOR_CONTRIBUYENTE_1	Factores que contribuyen a la colisión para el vehículo designado	Cadena
FACTOR_CONTRIBUYENTE_2	Factores que contribuyen a la colisión para el vehículo designado	Cadena

Colección y Descripción de los Datos

Nombre de la Columna	Descripción de la Columna	Tipo de Datos
ID_COLISIÓN	Código de registro único generado por el sistema	Cadena
FECHA_ACCIDENTE / FECHA_COLISIÓN	Fecha de ocurrencia de la colisión	Cadena
HORA_ACCIDENTE / HORA_COLISIÓN	Hora de ocurrencia de la colisión	Cadena
ID_ÚNICO	Código de registro único generado por el sistema	Cadena
ID_COLISIÓN	Código de identificación único del choque	Cadena
ID_VEHÍCULO	Código de identificación del vehículo asignado por el sistema	Cadena
TIPO_VEHICULO	Tipo de vehículo basado en la categoría seleccionada (ATV, bicicleta, automóvil/SUV, ebike, patinete eléctrico, camión/autobús, motocicleta, otro)	Cadena
MARCA_VEHÍCULO	Marca del vehículo	Cadena
MODELO_VEHÍCULO	Modelo del vehículo	Cadena
AÑO_VEHÍCULO	Año de fabricación del vehículo	Cadena
DIRECCIÓN_VIAJE	Dirección en la que se desplazaba el vehículo	Cadena
OCCUPANTES_VEHÍCULO	Número de ocupantes del vehículo	Cadena
SEXO_CONDUTOR	Género del conductor	Cadena
ESTADO_LICENCIA_CONDUTOR	Licencia, permiso, sin licencia	Cadena
JURISDICCIÓN_LICENCIA_CONDUTOR	Estado donde se emitió la licencia de conducir	Cadena
ACCIÓN_PRE_ACDNT	Ir recto, girar a la derecha, adelantar, retroceder, etc.	Cadena
PUNTO_DE_IMPACTO	Ubicación en el vehículo del punto de impacto oficial (es decir, lado del conductor, parte trasera del lado del pasajero, etc.)	Cadena
DANO_VEHÍCULO	Ubicación en el vehículo donde ocurrió la mayor parte del daño	Cadena
DANO_VEHÍCULO_1	Ubicaciones adicionales de daño en el vehículo	Cadena
DANO_VEHÍCULO_2	Ubicaciones adicionales de daño en el vehículo	Cadena
DANO_VEHÍCULO_3	Ubicaciones adicionales de daño en el vehículo	Cadena
DANO_PROPRIEDAD_PÚBLICA	Propiedad pública dañada (Sí o No)	Cadena
TIPO_DANO_PROPIEDAD_PÚBLICA	Tipo de propiedad pública dañada (por ejemplo, señal, cerca, poste de luz, etc.)	Cadena
FACTOR_CONTRIBUYENTE_1	Factores que contribuyen a la colisión para el vehículo designado	Cadena
FACTOR_CONTRIBUYENTE_2	Factores que contribuyen a la colisión para el vehículo designado	Cadena



Nombre de la Columna	Descripción de la Columna	Tipo de Datos
ID_COLISIÓN	Código de registro único generado por el sistema	Cadena
FECHA_ACCIDENTE / FECHA_COLISIÓN	Fecha de ocurrencia de la colisión	Cadena
HORA_ACCIDENTE / HORA_COLISIÓN	Hora de ocurrencia de la colisión	Cadena
ID_ÚNICO	Código de registro único generado por el sistema	Cadena
ID_COLISIÓN	Código de identificación único del choque	Cadena
ID_VEHÍCULO	Código de identificación del vehículo asignado por el sistema	Cadena
TIPO_VEHICULO	Tipo de vehículo basado en la categoría seleccionada (ATV, bicicleta, automóvil/SUV, ebike, patinete eléctrico, camión/autobús, motocicleta, otro)	Cadena
MARCA_VEHÍCULO	Marca del vehículo	Cadena
MODELO_VEHÍCULO	Modelo del vehículo	Cadena
AÑO_VEHÍCULO	Año de fabricación del vehículo	Cadena
DIRECCIÓN_VIAJE	Dirección en la que se desplazaba el vehículo	Cadena
OCCUPANTES_VEHÍCULO	Número de ocupantes del vehículo	Cadena
SEXO_CONDUTOR		

Colección y Descripción de los Datos

Nombre de la Columna	Descripción de la Columna	Tipo de Variable
CLAVE_DE_ARRESTO	ID persistente generado aleatoriamente para cada arresto	Cadena
FECHA_DE_ARRESTO	Fecha exacta del arresto del evento reportado	Fecha (sin hora)
CD_PD	Código de clasificación interno de tres dígitos (más detallado que el Código Clave)	Entero
DESC_PD	Descripción de la clasificación interna correspondiente con el código PD (más detallado que la descripción del delito)	Cadena
CD_KY	Código de clasificación interno de tres dígitos (categoría más general que el código PD)	Char
DESC_OFNS	Descripción de la clasificación interna correspondiente con el código KY (categoría más general que la descripción de PD)	Cadena
CODIGO_DELEY	Cargos de código de ley correspondientes a la Ley Penal del Estado de Nueva York, VTL y otras leyes locales varias	Cadena (puede ser separada)
CAT_CD_DELEY	Nivel de delito: felonía, delito menor, infracción	Cadena
BORO_DE_ARRESTO	Barrio del arresto. B (Bronx), S (Staten Island), K (Brooklyn), M (Manhattan), Q (Queens)	Char
PRECINTO_DE_ARRESTO	Precinto donde ocurrió el arresto	Entero

Nombre de la Columna	Descripción de la Columna	Tipo de Variable
CODIGO_DE_JURISDICCIÓN	Jurisdicción responsable del arresto. Los códigos de jurisdicción 0 (Patrulla), 1 (Tránsito) y 2 (Vivienda) representan al NYPD, mientras que los códigos 3 y más representan jurisdicciones no pertenecientes al NYPD	Entero
GRUPO_DE_EDAD	Edad del perpetrador dentro de una categoría	Cadena (puede ser cambiada)
SEXO_DEL_RPETRADOR	Descripción del sexo del perpetrador	Char
RAZA_DEL_RPETRADOR	Descripción de la raza del perpetrador	Cadena
COORD_X_CD	Coordenada X de bloque medio para el Sistema de Coordenadas del Estado de Nueva York, Zona de Long Island, NAD 83, unidades en pies (FIPS 3104)	Entero
COORD_Y_CD	Coordenada Y de bloque medio para el Sistema de Coordenadas del Estado de Nueva York, Zona de Long Island, NAD 83, unidades en pies (FIPS 3104)	Entero
Latitud	Coordenada de latitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales (EPSG 4326)	Flotante
Longitud	Coordenada de longitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales (EPSG 4326)	Flotante

Colección y Descripción de los Datos

Nombre de la Columna	Descripción de la Columna	Tipo de Variable
CLAVE_DE_AR	ID persistente generado aleatoriamente para cada arresto	Cadena
RESTO	Fecha exacta del arresto del evento reportado	Fecha (sin hora)
CD_KY	Código de clasificación interno de tres dígitos (más detallado que el Código Clave)	Entero
DESC_PD	Descripción de la clasificación interna correspondiente con el código PD (más detallado que la descripción del delito)	Cadena
CODIGO_DE_L_EY	Código de código de ley correspondientes a la Ley Penal del Estado de Nueva York, VTL y otras leyes locales varias	Cadena (puede ser separada)
CAT_CD_DE_L_EY	Barrio del arresto. A (Bronx), B (Bronx), S (Staten Island), K (Brooklyn), M (Manhattan), Q (Queens)	Cadena
BORO_DE_AR	Char	
PRECINTO_DE_ARRESTO	Precinto donde ocurrió el arresto	Entero
CODIGO_DE_JURISDICCIÓN	Jurisdicción responsable del arresto. Los códigos de jurisdicción 0 (Patrulla), 1 (Tránsito) y 2 (Vivienda) representan al NYPD, mientras que los códigos 3 y más representan jurisdicciones no pertenecientes al NYPD	Entero
GRUPO_DE_EDAD	Edad del perpetrador dentro de una categoría	Cadena (puede ser cambiada)
SEXO_DEL_PEPETRADOR	Descripción del sexo del perpetrador	Char
RAZA_DEL_PEPETRADOR	Descripción de la raza del perpetrador	Cadena
COORD_X_CD	Coordenada X de bloque medio para el Sistema de Coordenadas del Estado de Nueva York, Zona de Long Island, NAD 83, unidades en pies (FIPS 3104)	Entero
COORD_Y_CD	Coordenada Y de bloque medio para el Sistema de Coordenadas del Estado de Nueva York, Zona de Long Island, NAD 83, unidades en pies (FIPS 3104)	Entero
Latitud	Coordenada de latitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales (EPSG 4326)	Flotante
Longitud	Coordenada de longitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales (EPSG 4326)	Flotante



Nombre de la Columna	Descripción de la Columna	Tipo de Variable
CLAVE_DE_AR	ID persistente generado aleatoriamente para cada arresto	Cadena
FECHA_DE_AR	Fecha exacta del arresto del evento reportado	Cadena
CD_PD	Código de clasificación interno de tres dígitos (más detallado que el Código Clave)	Cadena
DESC_PD	Descripción de la clasificación interna correspondiente con el código PD (más detallado que la descripción del delito)	Cadena
CD_KY	Código de clasificación interno de tres dígitos (categoria más general que el código PD)	Cadena

03

Análisis Exploratorio

Estadística descriptiva para el entendimiento de los Datos

¿Qué se hizo?

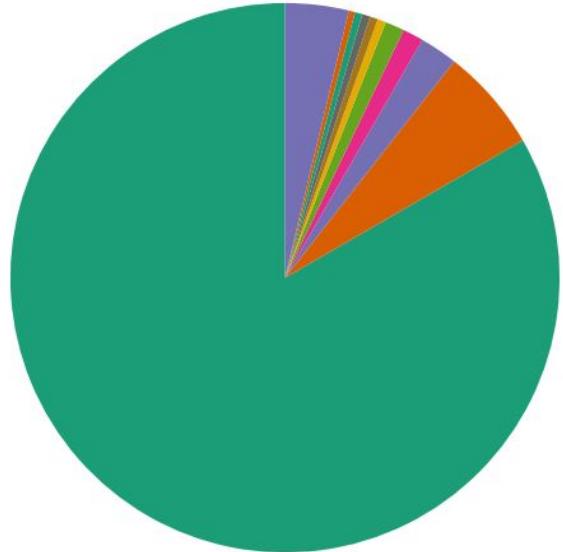
Se hizo uso de la librería matplotlib para graficar ciertas columnas de interés que pudieran dar ciertos indicios para los análisis predictivos así como para tener un entendimiento general de los datos.



EDA Colisiones

EDA Colisiones

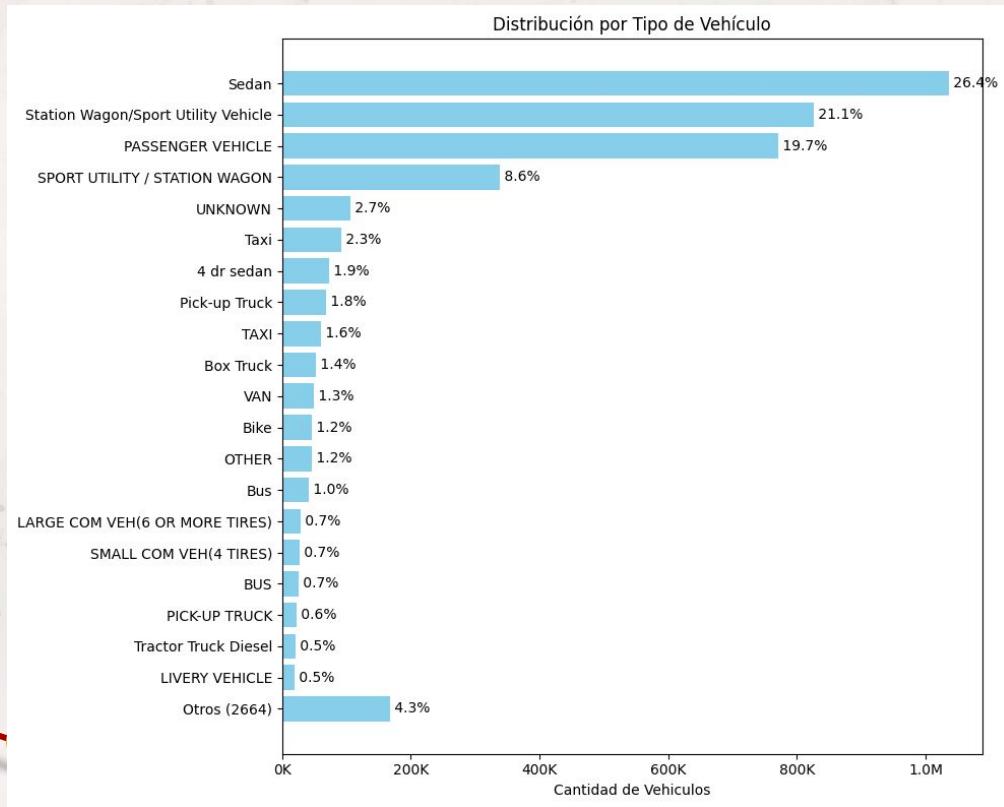
Distribución de Registro de Vehículos por Estado



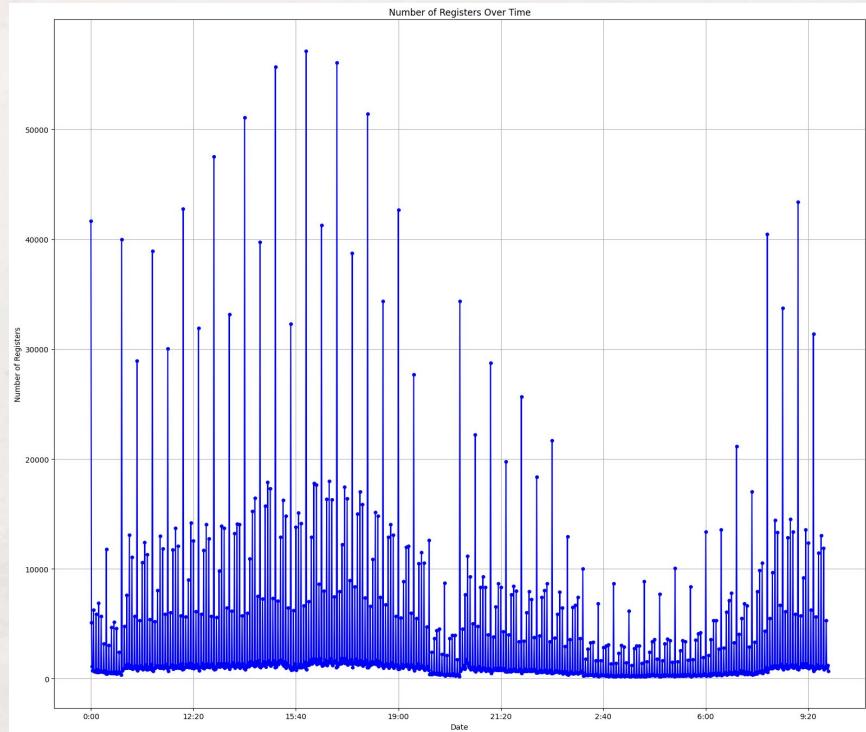
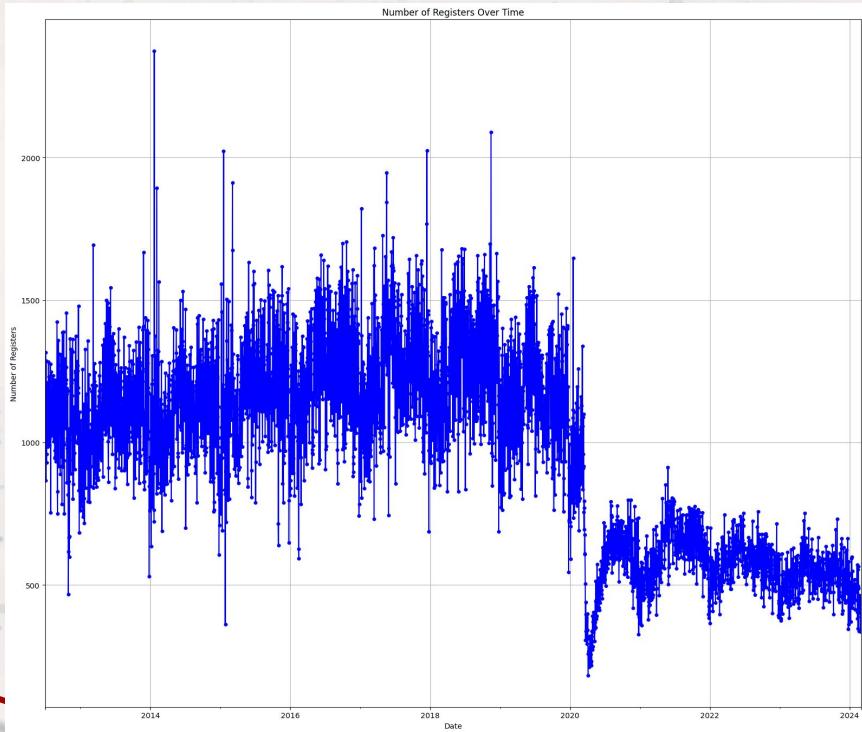
- New York (NY): 83.3%
- New Jersey (NJ): 6.1%
- Pennsylvania (PA): 2.2%
- Florida (FL): 1.2%
- Connecticut (CT): 1.1%
- Virginia (VA): 0.5%
- Massachusetts (MA): 0.5%
- Maryland (MD): 0.5%
- North Carolina (NC): 0.4%
- Texas (TX): 0.4%
- Otros (72) (Otros (72)): 3.7%



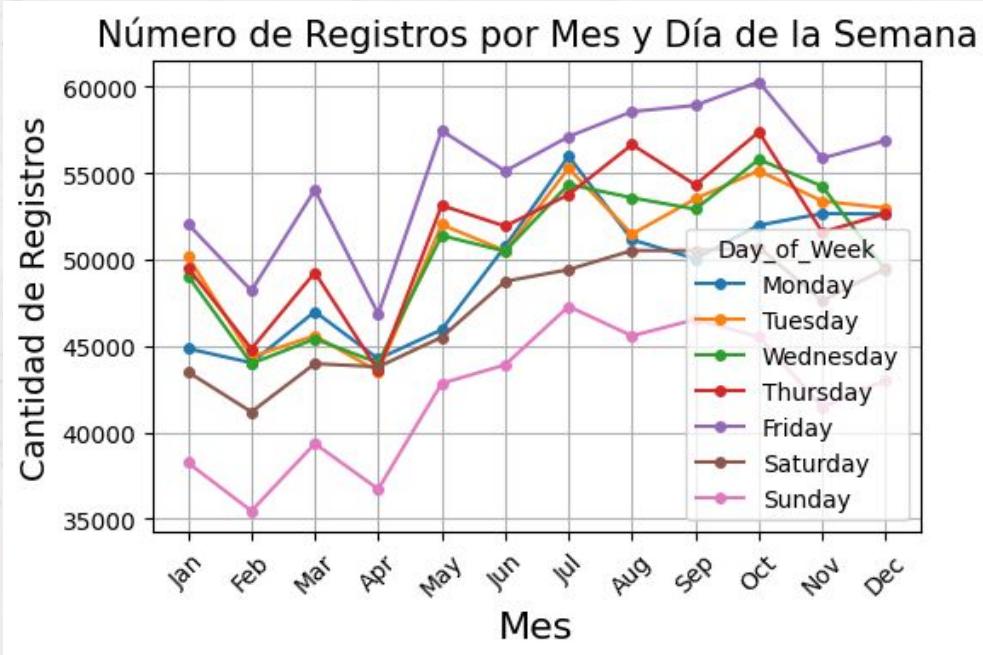
EDA Colisiones



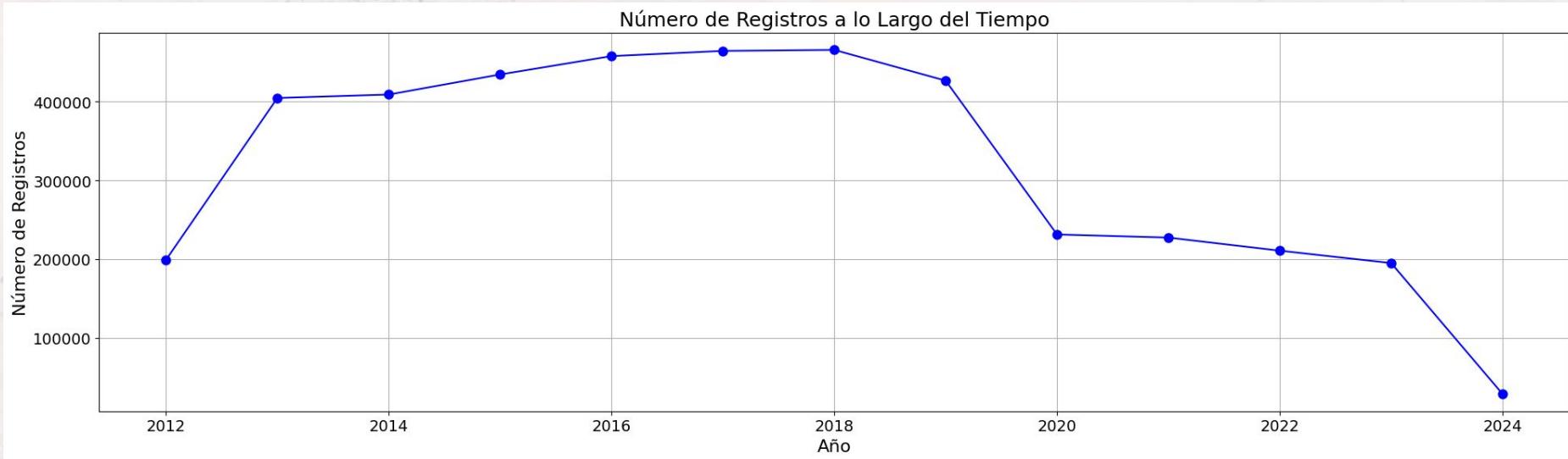
EDA Colisiones



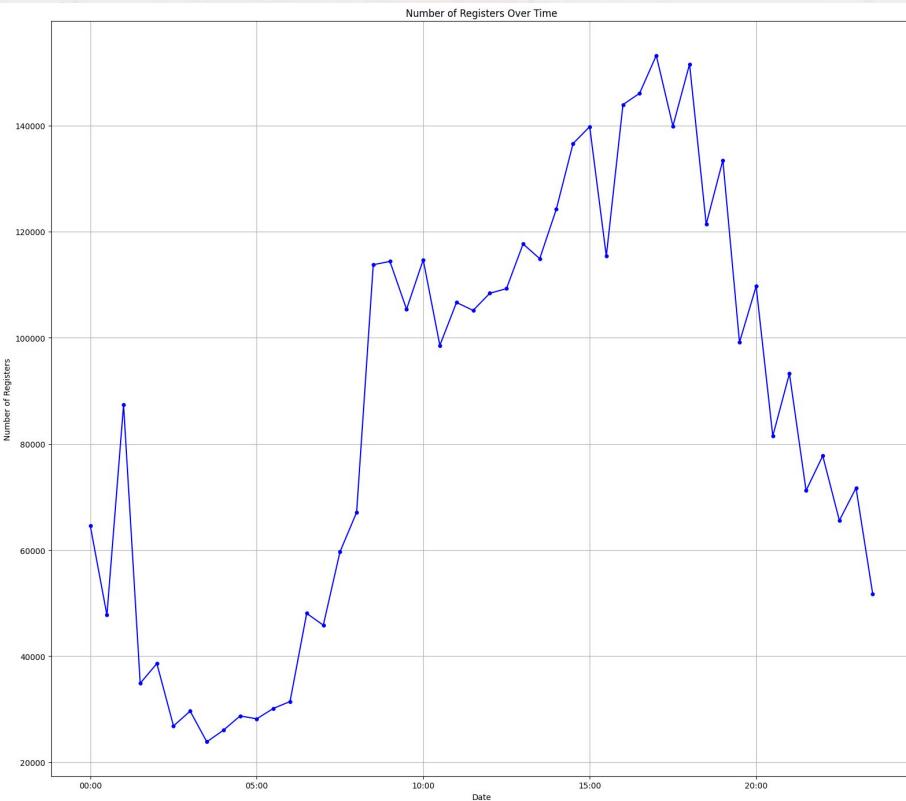
EDA Colisiones



EDA Colisiones

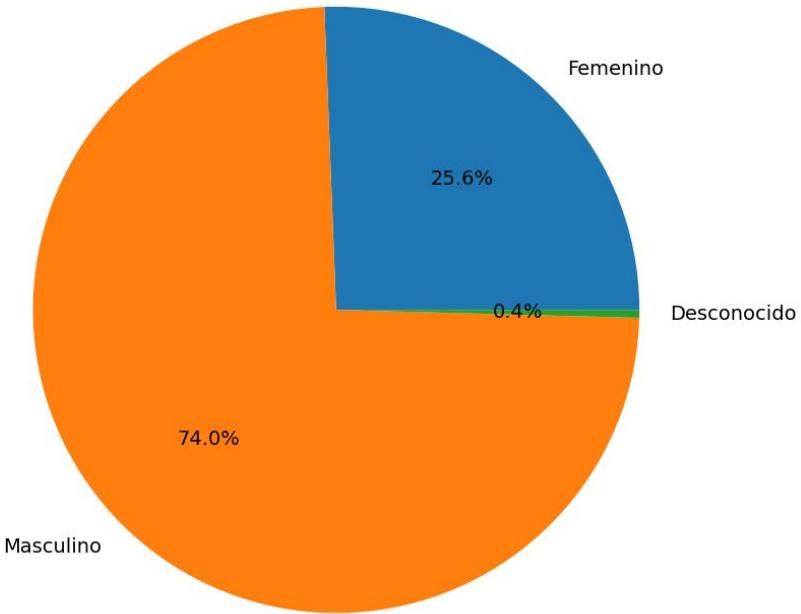


EDA Colisiones



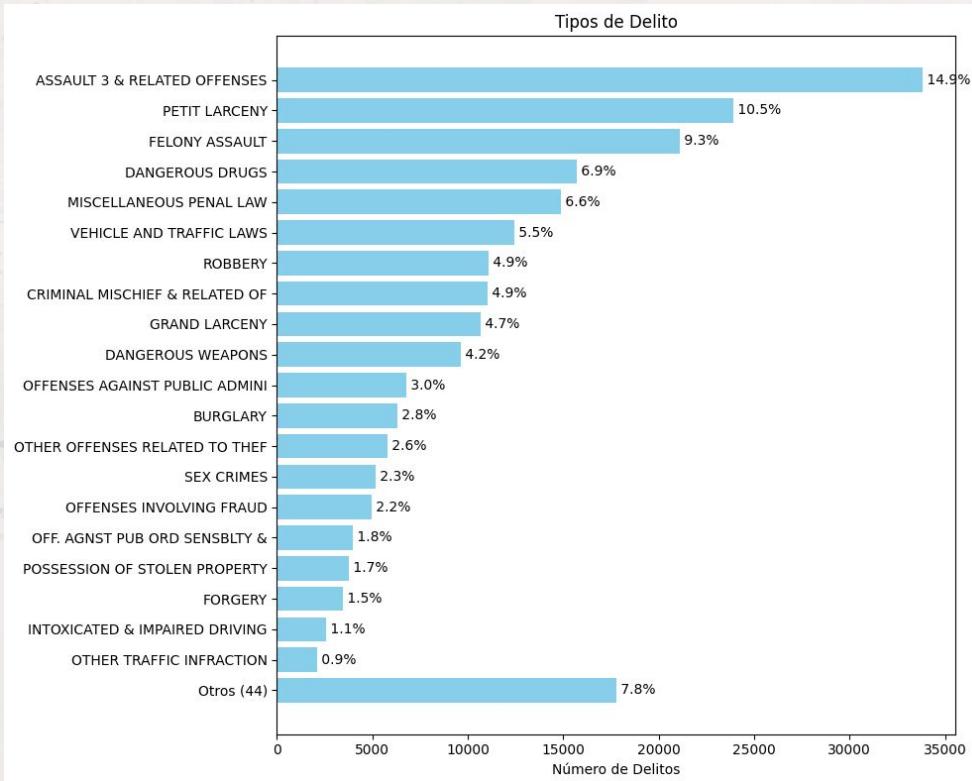
EDA Colisiones

Sexo del conductor en el choque

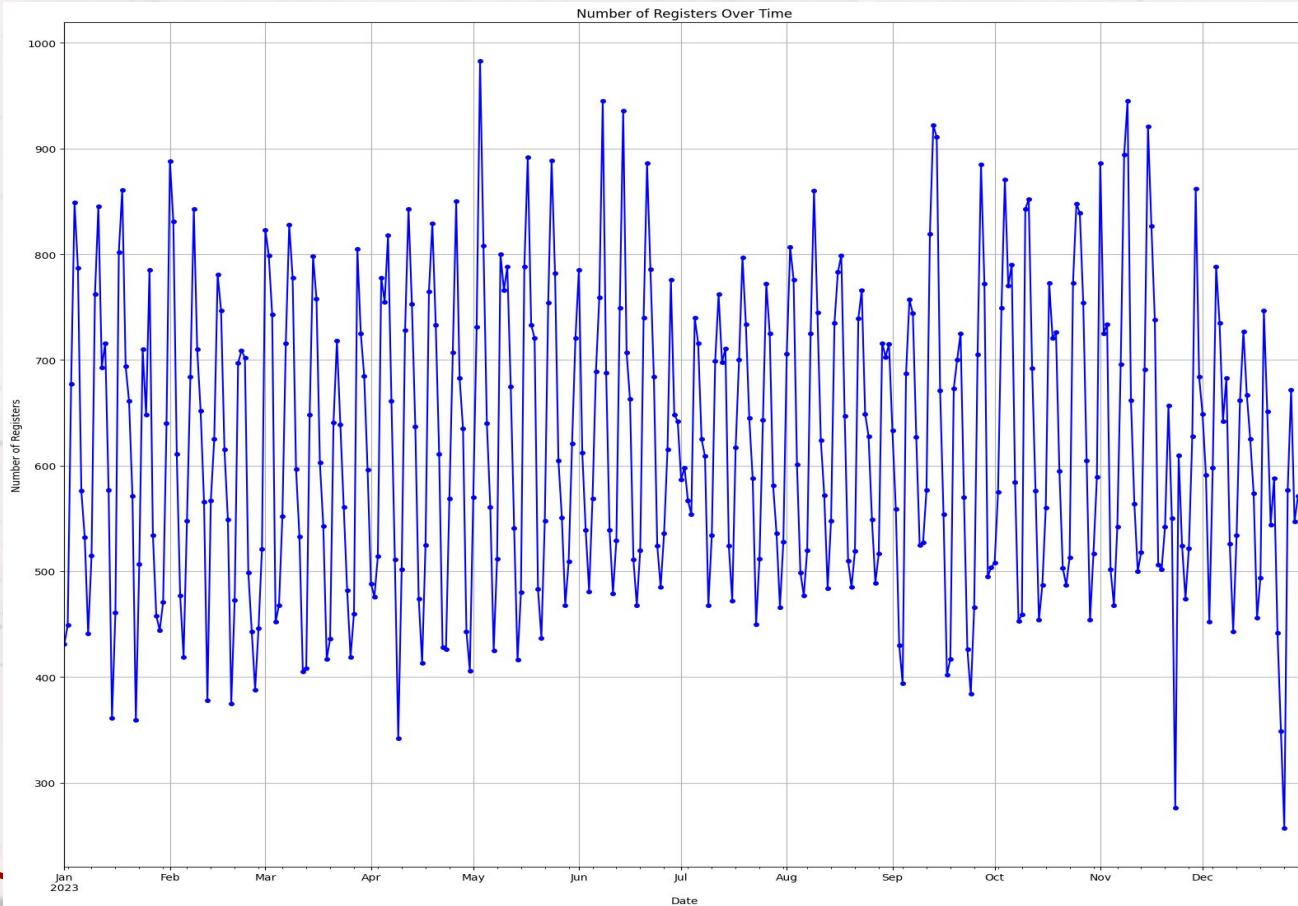


EDA Arrestos

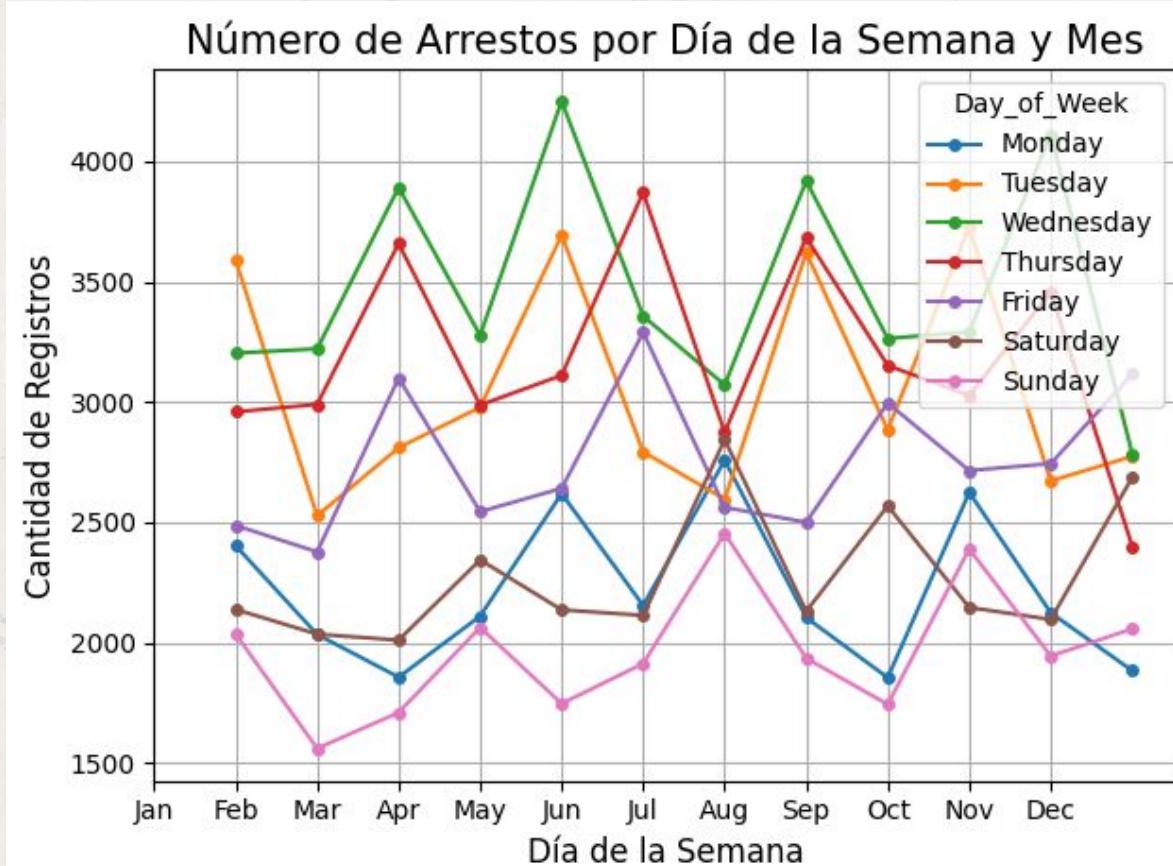
EDA Arrestos



EDA Arrestos

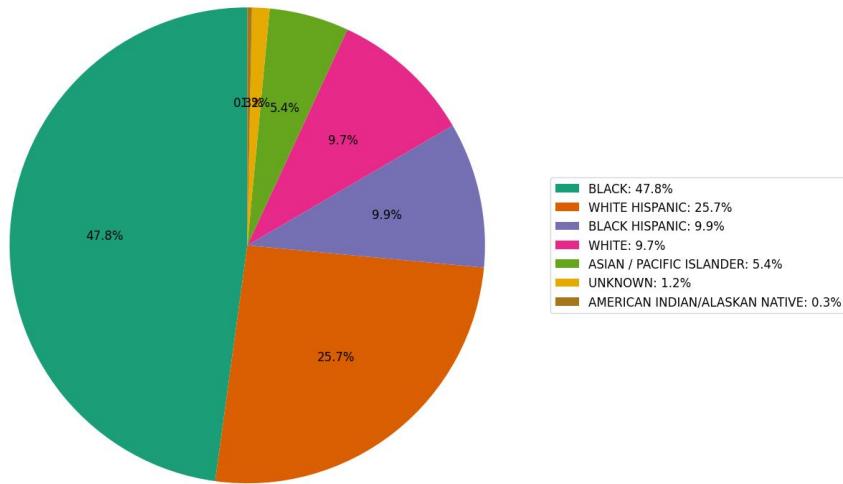


EDA Arrestos

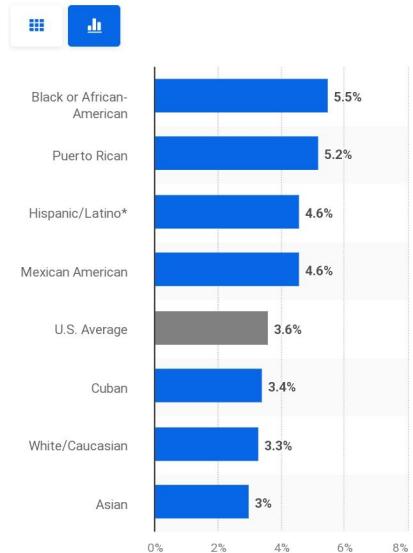


EDA Arrestos

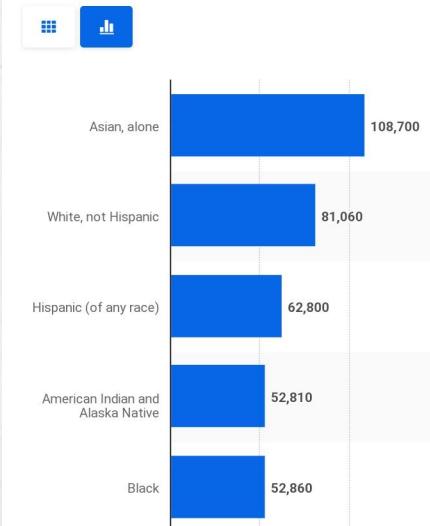
Proporción de Arrestos por Raza



Unemployment rate in the United States in 2023, by ethnicity

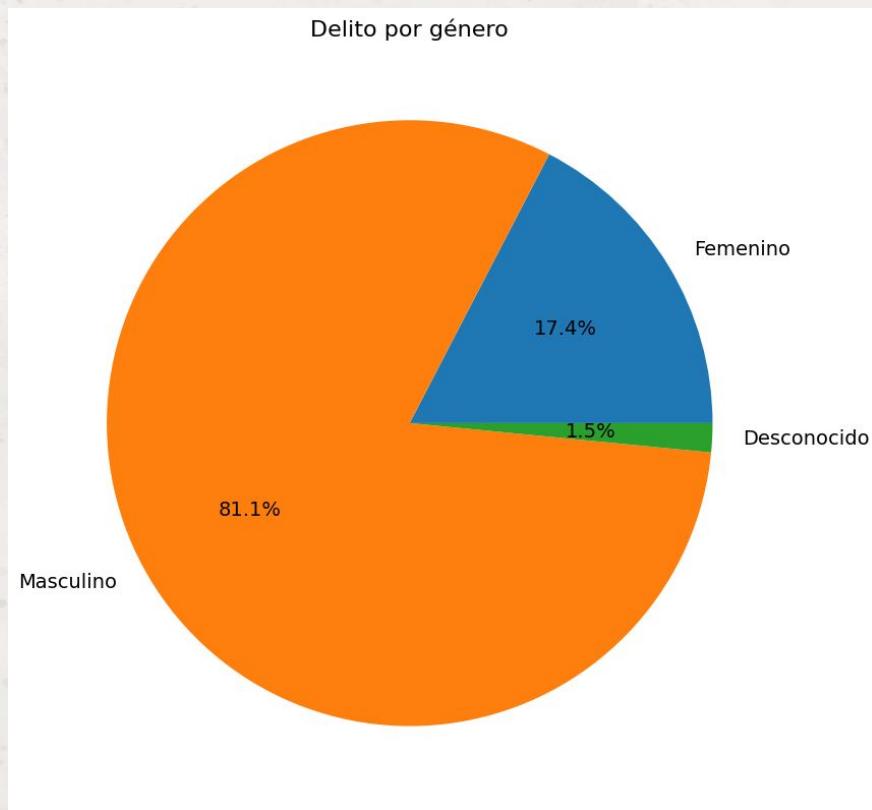


Median household income in the United States in 2022, by race and ethnicity
(in U.S. dollars)



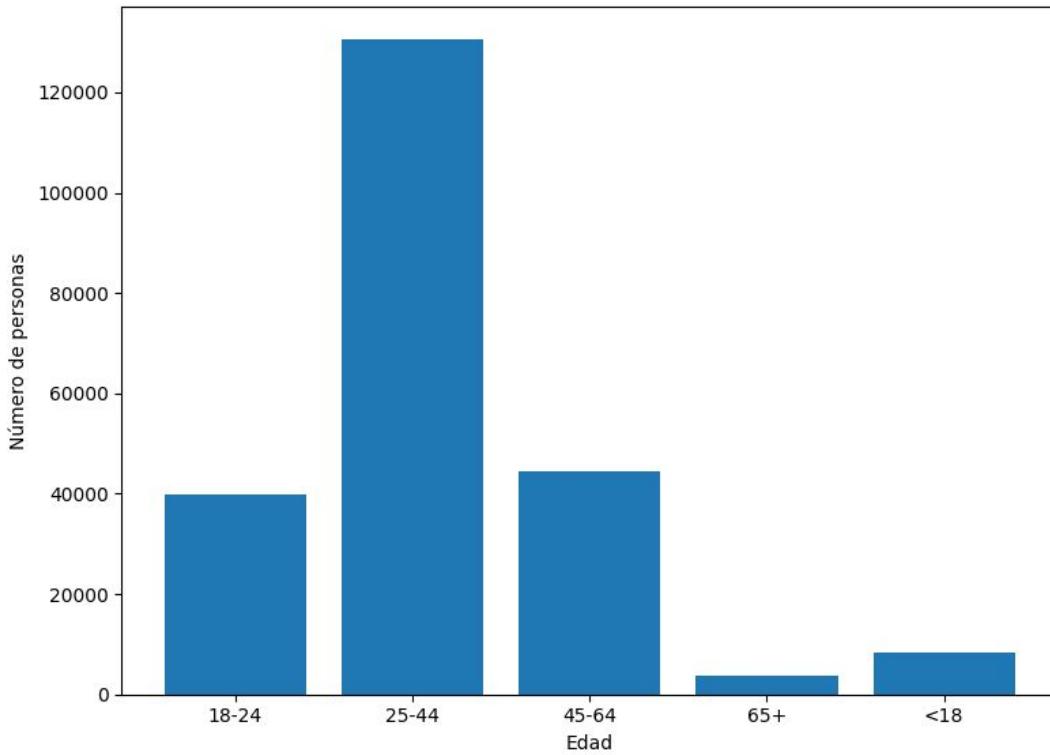
<https://www.statista.com/statistics/233324/median-household-income-in-the-united-states-by-race-or-ethnic-group/>
<https://www.statista.com/statistics/237917/us-unemployment-rate-by-race-and-ethnicity/>

EDA Arrestos



EDA Arrestos

Cantidad de delitos cometidos basados en la edad



04

Reporte de Calidad

Cantidad de valores nulos y acciones a tomar

DataSet Arrestos

Nombre Columna	Número de Nulos
ARREST_KEY	0
ARREST_DATE	0
PD_CD	2
PD_DESC	0
OFNS_DESC	17
LAW_CODE	0
LAW_CAT_CD	0
ARREST_BORO	1599
ARREST_PRECI	0
JURISDICTION_	0
AGE_GROUP	0
PERP_SEX	0
PERP_RACE	0
X_COORD_CD	0
Y_COORD_CD	0
Latitude	0
Longitude	0
New Georereduced	0

DataSet Arrestos



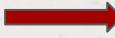
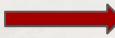
Nombre Columna	Número de Nulos
ARREST_KEY	0
ARREST_DATE	0
PD_CD	2
PD_DESC	0
OFNS_DESC	17
LAW_CODE	0
LAW_CAT_CD	0
ARREST_BORO	1599
ARREST_PRECI	0
JURISDICTION_	0
AGE_GROUP	0
PERP_SEX	0
PERP_RACE	0
X_COORD_CD	0
Y_COORD_CD	0
Latitude	0
Longitude	0
New Georereneced	0

DataSet Colisiones

Nombre Columna	Número de Nulos
UNIQUE_ID	0
COLLISION_ID	0
CRASH_DATE	0
CRASH_TIME	0
VEHICLE_ID	0
STATE_REGISTRATION	299985
VEHICLE_TYPE	233571
VEHICLE_MAKE	1875745
VEHICLE_MODEL	4103370
VEHICLE_YEAR	1894945
TRAVEL_DIRECTION	1665989
VEHICLE_OCCUPANTS	1778904
DRIVER_SEX	2210888
DRIVER_LICENSE_STATUS	2298728
DRIVER_LICENSE_JURISDICTION	2293835
PRE_CRASH	919020
POINT_OF_IMPACT	1698852
VEHICLE_DAMAGE	1722871
VEHICLE_DAMAGE_1	2589444
VEHICLE_DAMAGE_2	2976594
VEHICLE_DAMAGE_3	3252319
PUBLIC_PROPERTY_DAMAGE	1528858
PUBLIC_PROPERTY_DAMAGE_TYPE	4128961
CONTRIBUTING_FACTOR_1	146425
CONTRIBUTING_FACTOR_2	1685712

DataSet Colisiones

Nombre Columna	Número de Nulos
UNIQUE_ID	0
COLLISION_ID	0
CRASH_DATE	0
CRASH_TIME	0
VEHICLE_ID	0
STATE_REGISTRATION	299985
VEHICLE_TYPE	233571
VEHICLE_MAKE	1875745
VEHICLE_MODEL	4103370
VEHICLE_YEAR	1894945
TRAVEL_DIRECTION	1665989
VEHICLE_OCCUPANTS	1778904
DRIVER_SEX	2210888
DRIVER_LICENSE_STATUS	2298728
DRIVER_LICENSE_JURISDICTION	2293835
PRE_CRASH	919020
POINT_OF_IMPACT	1698852
VEHICLE_DAMAGE	1722871
VEHICLE_DAMAGE_1	2589444
VEHICLE_DAMAGE_2	2976594
VEHICLE_DAMAGE_3	3252319
PUBLIC_PROPERTY_DAMAGE	1528858
PUBLIC_PROPERTY_DAMAGE_TYPE	4128961
CONTRIBUTING_FACTOR_1	146425
CONTRIBUTING_FACTOR_2	1685712



05

Planteamiento de Preguntas

Preguntas relacionada

DataSet Colisiones



Horas vs Accidentalidad

¿Cuáles son las horas o intervalos de tiempo en dónde se presenta mayor accidentalidad (y en qué zonas)?



Tipo de Daño

¿ Se podría predecir el tipo de daño o la probabilidad de algún tipo de daño basado en algún perfil?



Top 3 vehículos

¿ Cuáles son el top 3 de vehículos más propensos a generar accidentes?



Perfil

¿ Hay algún perfil (conjunto de características) en específico que pueda generar mayores niveles de accidentalidad?

Statistical interpretation



Perfil de Zona y Crimen

¿ Se podría, basado en un perfil y zona, predecir el crimen ocurrido?



Perfilamiento

¿ Hay algún perfil/clasificación para un crimen específico?



Geográfico

¿ En qué zonas se comete que cantidad de crímenes con más frecuencia?



Correlación

¿Qué correlación raza, género, edad y el tipo de crimen?

06

Filtro, limpieza y transformación

Estandarización, eliminación de columnas innecesarias
entre otros

DataSet Arrestos

```
1 #se va eliminar la columna de KY_CD porque PD_CD ya contiene esos datos en otro formato
2
3 df_filled = df_filled.drop("KY_CD")
4
5 # También se van a eliminar las columnas ARREST_KEY, LAW_CODE, JURISDICTION_CODE, New Georeferenced
6 # Column y LAW_CAT_CD porque no aporta nada para el ejercicio de análisis
7
8 df_filled = df_filled.drop("ARREST_KEY")
9 df_filled = df_filled.drop("LAW_CODE")
10 df_filled = df_filled.drop("JURISDICTION_CODE")
11 df_filled = df_filled.drop('New Georeferenced Column')
12 df_filled = df_filled.drop('LAW_CAT_CD')
13
14 df_filled.limit(4).toPandas()
15
16
```

DataSet Colisiones

```
1 from pyspark.sql.functions import split  
2  
3 # Split the VEHICLE_MAKE  
4 df = df.withColumn("MAKE", split(df["VEHICLE_MAKE"], "-")[0]) \  
5     .withColumn("MODEL", split(df["VEHICLE_MAKE"], "-")[1])  
6  
7 # Drop VEHICLE_MAKE |  
8 df = df.drop("VEHICLE_MAKE")  
9
```

```
1 #Se elimina la columna model porque la columnas Vehicle type aporta la misma información con menor  
2 cantidad de nulos  
3 df = df.drop('MODEL')  
4 # Se van a eliminar las siguientes columnas debido a que no se van a usar para el análisis  
5 df = df.drop('UNIQUE_ID')  
6 df = df.drop('VEHICLE_ID')  
7 df = df.drop('PUBLIC_PROPERTY_DAMAGE')  
8 df = df.drop('PUBLIC_PROPERTY_DAMAGE_TYPE')  
9 df = df.drop('VEHICLE_MODEL')
```

DataSet Colisiones

```
1  def round_time(time):
2      hour, minute = map(int, time.split(':'))
3      if 15 < minute <= 45:
4          minute = 30
5      else:
6          minute = 0
7          if minute == 0 and hour == 23:
8              hour = 0
9          else:
10             hour = (hour + 1) if minute == 0 else hour
11      return f"{hour:02d}:{minute:02d}"
12
13  print(round_time('10:29'))
14
15 # Register UDF
16 round_time_udf = udf(round_time, StringType())
17
18 # Apply UDF to CRASH_TIME column
19 df = df.withColumn("CRASH_TIME", round_time_udf("CRASH_TIME"))
20 df.limit(4).toPandas()
21
```

CRASH_TIME	S
	9:00
	8:00
	17:00
	20:30

DataSet Colisiones

```
1  from pyspark.sql.functions import when, col
2
3  # List of possible damage types
4  damage_types = [
5      "Left Rear Quarter Panel", "Center Back End", "Left Front Quarter Panel",
6      "None", "Right Front Bumper", "Left Rear Bumper", "Right Side Doors",
7      "Roof", "Right Front Quarter Panel", "Other", "Overturned", "Trailer",
8      "Right Rear Quarter Panel", "Left Side Doors", "Left Front Bumper",
9      "Demolished", "Center Front End", "Undercarriage", "Right Rear Bumper",
10     "No Damage"
11 ]
12
13 # Loop through each damage type and apply when condition to fill respective column
14 for damage_type in damage_types:
15     column_name = damage_type.replace(" ", "_") # Convert damage type to column name format
16     # Check if the column already has a value of 1, if so, retain it, otherwise, set based on
17     # VEHICLE_DAMAGE_1
18     df = df.withColumn(column_name,
19                         when((col(column_name) == 1) | (df["VEHICLE_DAMAGE"] == damage_type), 1)
20                         .otherwise(0))
21
```

DataSet Colisiones

POINT_OF_IMPACT	VEHICLE_DAMAGE	VEHICLE_DAMAGE_1	VEHICLE_DAMAGE_2	VEHICLE_DAMAGE_3	CONTRIBUTOR
Left Front Bumper	Left Front Quarter Panel	None	None	None	Driver Inattentive
Right Front Bumper	Right Front Bumper	Right Front Quarter Panel	None	None	Driver Inattentive
Left Front Quarter Panel	Left Front Quarter Panel	None	None	None	Driver Inattentive
Center Front End	Center Front End	No Damage	No Damage	No Damage	Driver Inattentive
Right Rear Bumper	Right Rear Bumper	Center Back End	Left Rear Bumper	None	Driver Inattentive
...
Right Front Bumper	Right Front Bumper	No Damage	No Damage	No Damage	Driver Inattentive
Center Front End	Center Front End	None	None	None	Driver Inattentive
No Damage	No Damage	No Damage	No Damage	No Damage	Pedestrian Pedestrian
Right Front Bumper	Right Front Bumper	Right Front Quarter Panel	Center Back End	None	Driver Inattentive
Center Back End	Center Back End	Left Rear Bumper	Right Rear Bumper	None	Driver Inattentive

Panel	Center_Back_End	Left_FrontQuarter_Panel	None	Right_Front_Bumper	Left_Rear_Bumper	Right_Side_Doors	Ro
0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
0	1	0	0	0	1	0	0
...
0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0
0	1	0	0	0	0	1	0

Gracias !

me after
watching fake
horror movies:



me after
watching true
serial killer
documentaries:

Take that criminals



me after
watching fake
horror movies:
me after
watching true
serial killer
documentaries:

me after
watching fake
horror movies:
me after
watching true
serial killer
documentaries:

me after
watching fake
horror movies:
me after
watching true
serial killer
documentaries: