Primera Entrega de Proyecto: Procesamiento de Datos a Gran Escala

Isaac Janica, Daniela Torres,Daniel Sandoval

[1]Ciencia de Datos& Pontifica Universidad Javeriana, Colombia

[2]Ingeniería de Sistemas & Pontifica Universidad Javeriana, Colombia

[1]daniela.torresg@javeriana.edu.co; janica.i@javeriana.edu.co;daniel_sandoval@javeriana.edu.co

Abstract— ***This document contains all the documentation with respect to the datasets of Collisions and Arrest that were chosen for the Big Data class Project.***

Keywords— DataSet, Collision, PIB, Arrest.

I. Introduction

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

II. Entendimiento del negocio

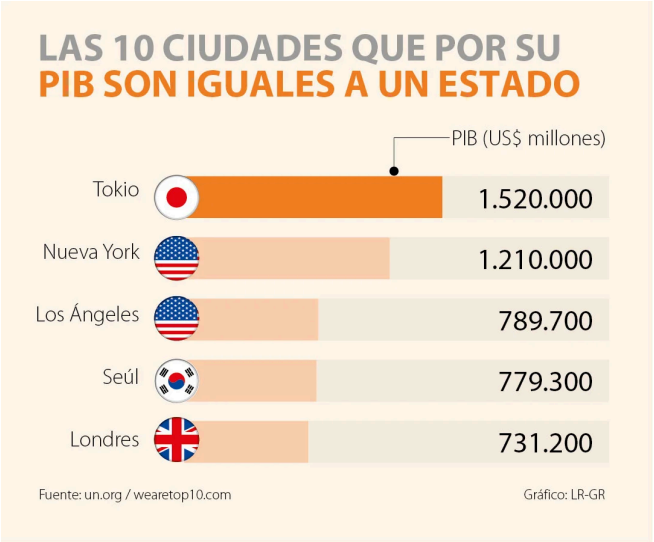
Si bien puede parecer al lector un poco fuera de lugar este inicio, se va a empezar este apartado hablando un poco del actual alcalde de Nueva York. Eric Adams antes de empezar la carrera política, fue un oficial del NYPD (New York Police Department), después de esto fue Senador, presidente del Borough del Bronx. Las propuestas que lo llevaron a llegar al cargo que actualmente ocupa es: "traer de vuelta" la economía de NY, reducir la desigualdad, mejor seguridad de públic (public safety) y construir una ciudad más fuerte y saludable para todos los Neoyorkinos.

En este sentido, uno de los apartados al que más se le va a hacer énfasis es el de la seguridad, debido a que el Alcalde buscaría poder tener estos indicadores altos por lo que él fue un miembro de la policía, así cómo en sus propuestas prometió mejor la seguridad dentro de la ciudad.

Respecto a ciertos indicadores macroeconómicos, uno de los que más llama la atención es que New York es el estado que cuenta con el mayor PIB per cápita de los Estados Unidos como se puede observar en la siguiente gráfica:

Esto ilustra el hecho de que en esta ciudad los bienes que portan y poseen los Neoyorkinos puede tener un valor más alto que el resto de los estados, es decir, que si se quisiera hurtar a una persona de Estados Unidos, una de las personas más rentables, sería una persona de Nueva York.

Esta cifra puede ser complementada con el hecho de que NYC maneja cantidades de dinero tan grandes que su PIB es 4 veces el PIB actual de Colombia (314.5 millones de USD). Esto se puede observar en los siguientes gráfico:



A. Indicadores generales de Seguridad

Para las estadísticas y la situación general del Estado tenemos que 5 de los 7 indicadores principales que tiene el departamento de Policía, las cifras tuvieron un decremento: en asesinatos hubo una reducción del 25%, las violaciones un 24.4%, el delito grave de asalto un 1.5%, el allanamiento de morada se redujo un 19.8% así cómo el hurto mayor de automóviles un 3.8%.

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%**

Sin embargo, no todos los resultado son positivos, en el caso del Robo/atraco hubo un aumento del 5.4% así cómo el hurto mayor.

Ahora, una de las estadísticas que actualmente preocupan a la ciudadanía es el hecho del aumento de los crímenes de odio, específicamente, odio hacia personas judías, esto se puede observar en la siguiente gráfica:

Hate Crimes Statistics: January 2024

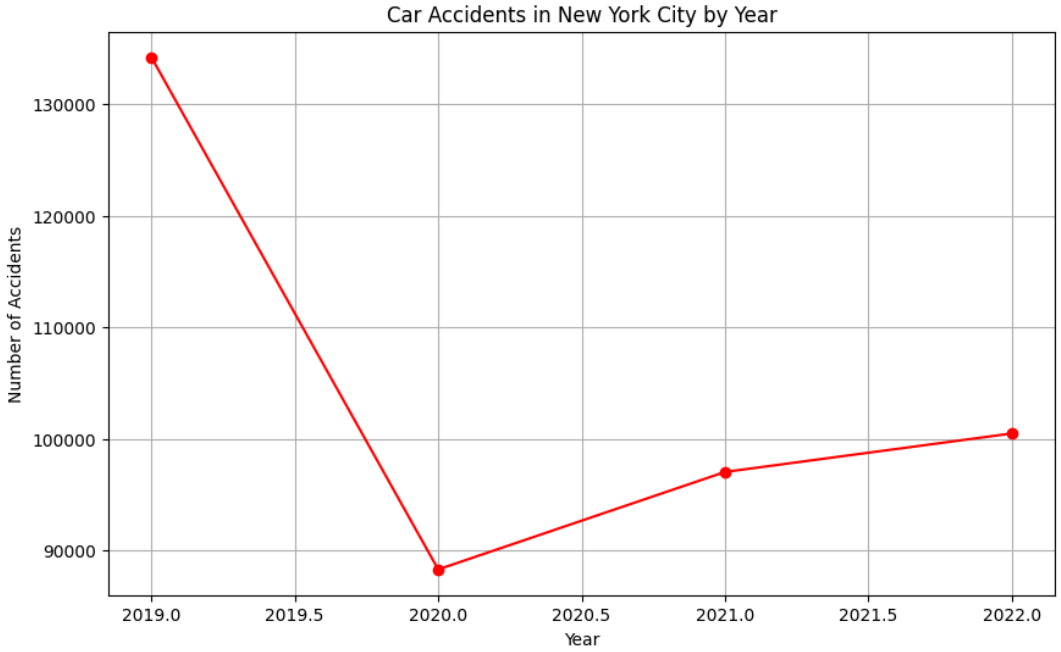
(Representing January 1 – January 31 for calendar years 2024 and 2023)

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%**

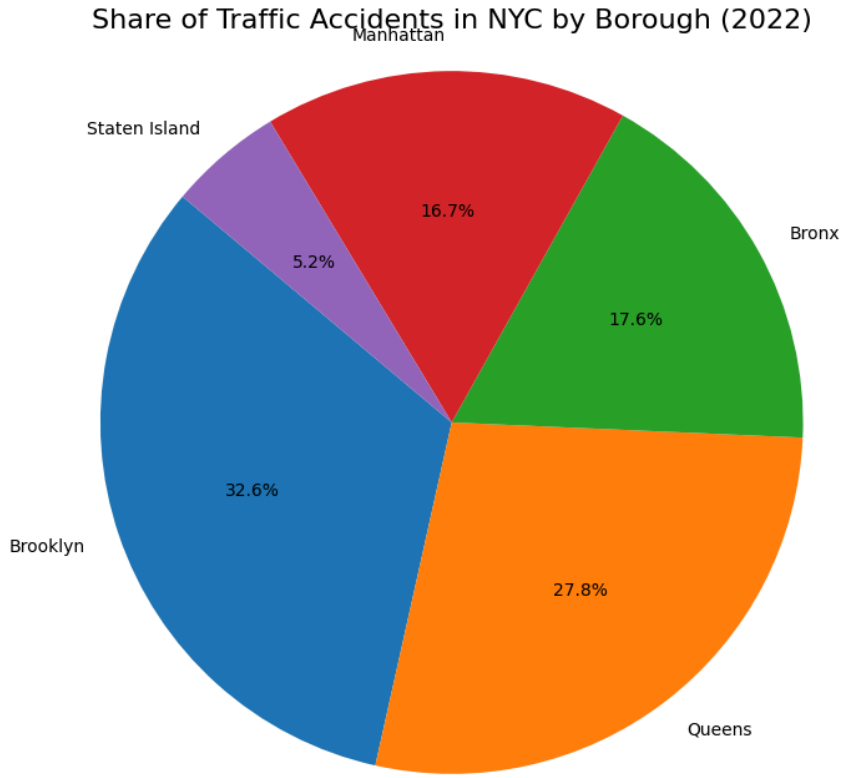
Cómo se puede observar hubo un aumento del 82% en crímenes de odio contra las personas Judías, esto fue uno de los mayores causantes de un incremento del 36% de crímenes de odio.

Crímenes por ubicación:
https://propertyclub.nyc/article/most-dangerous-neighborhoods-in-nyc

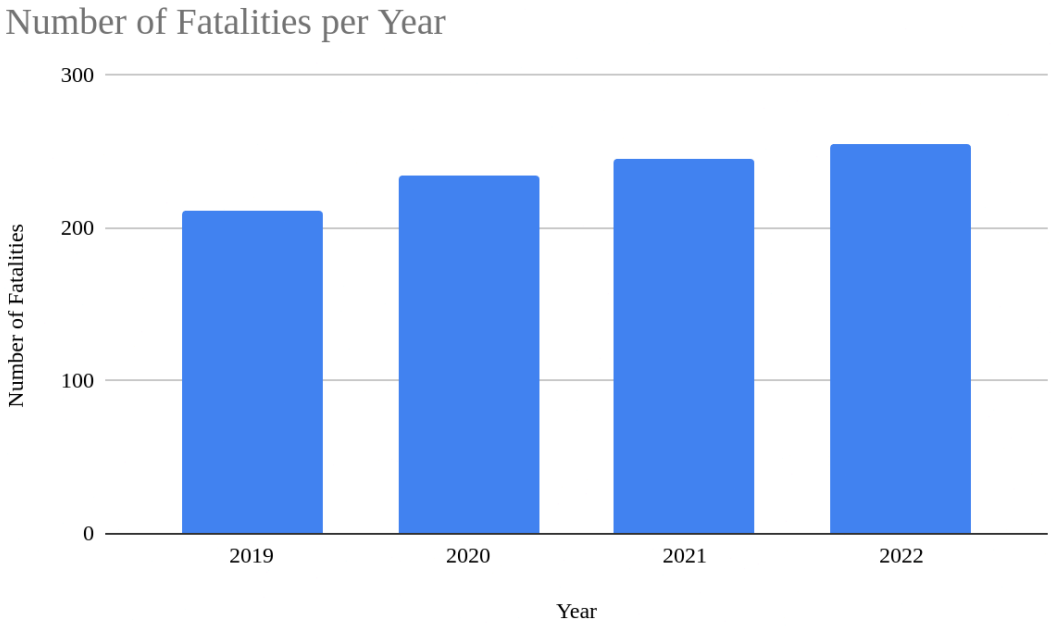
Respecto a los accidentes en la ciudad de New York se halló que en 2019 hubo un total de 134,224 accidentes, sin embargo, en 2020 esté número decreció a raíz de la pandemia del covid-19 a 88,323. Ahora, en cifras más actuales, se encontró que hubo un total de 100,508 accidentes viales en 2022, un aumento a comparación de 2020 pero no a comparación de 2019.



También se encontró que las 'localidades' que tienen mayor accidentalidad son Brooklyn y Queens, juntas contienen el 60% de los accidentes dentro de la ciudad.



Ahora, con respecto a la cantidad de accidentes que terminan muertes ha tenido un crecimiento sostenido durante 2019 a 2022, esto se puede observar en el siguiente gráfico:



Las tres localidades con mayor número de muertes por estos accidentes son Brooklyn (32%), Queens (24.9%) y Bronx con (19.8%). Ahora, el tiempo/fechas en dónde pueden ocurrir estos accidentes varía, en 2021 Brooklyn tenía la mayoría de los incidentes/mes en Septiembre con 2850, Manhattan tenía la mayor cantidad en octubre con 1482. Además, la causa que mayor causó colisiones es la 'distracción al manejar' la cuál tiene el 29.8% de las colisione, esta es seguida de otras cómo estar muy cerca al carro vecino o un cruce inapropiado.

III. Colección y descripción de los datos

Para este apartado se prefirió hacer una tabla que recolectara el nombre de la columna, la descripción del dato de la columna y el tipo de variable **que tiene** la columna, es decir, que la columna en el dataset original podía ser de tipo object, pero por motivos de estandarización, se convirtió esta a tipo string. Con esto en mente, esta es la tabla/Diccionario de Datos del DataSet de arrestos:

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	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 (categoría más general que el código PD)	Cadena

DESC_OFNS	Descripción de la clasificación interna correspondiente con el código KY (categoría más general que la descripción de PD)	Cadena
CODIGO_DE_LEY	Cargos de código de ley correspondientes a la Ley Penal del Estado de Nueva York, VTL y otras leyes locales varias	Cadena
CAT_CD_DE_LEY	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)	Cadena
PRECINTO_DE_ARRESTO	Precinto donde ocurrió el arresto	Cadena
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	Cadena
GRUPO_DE_EDAD	Edad del perpetrador dentro de una categoría	Cadena
SEXO_DEL_PERPETRADOR	Descripción del sexo del perpetrador	Cadena
RAZA_DEL_PERPETRADOR	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)	Cadena
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)	Cadena
Latitud	Coordenada de latitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales (EPSG 4326)	Cadena
Longitud	Coordenada de longitud para el Sistema de Coordenadas Globales, WGS 1984, grados decimales (EPSG 4326)	Cadena

Vale la pena resaltar que los nombres de las columnas fueron traducidos al español. Ahora, la siguiente tabla es el Diccionario de Datos el DataSet de colisiones:

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
OCUPANTES_VEHÍCULO	Número de ocupantes del vehículo	Cadena
SEXO_CONDUCTOR	Género del conductor	Cadena
ESTADO_LICENCIA_CONDUCTOR	Licencia, permiso, sin licencia	Cadena
JURISDICCIÓN_LICENCIA_CONDUCTOR	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_PROPIEDAD_PÚBLICA	Propiedad pública dañada (Sí o No)	Cadena

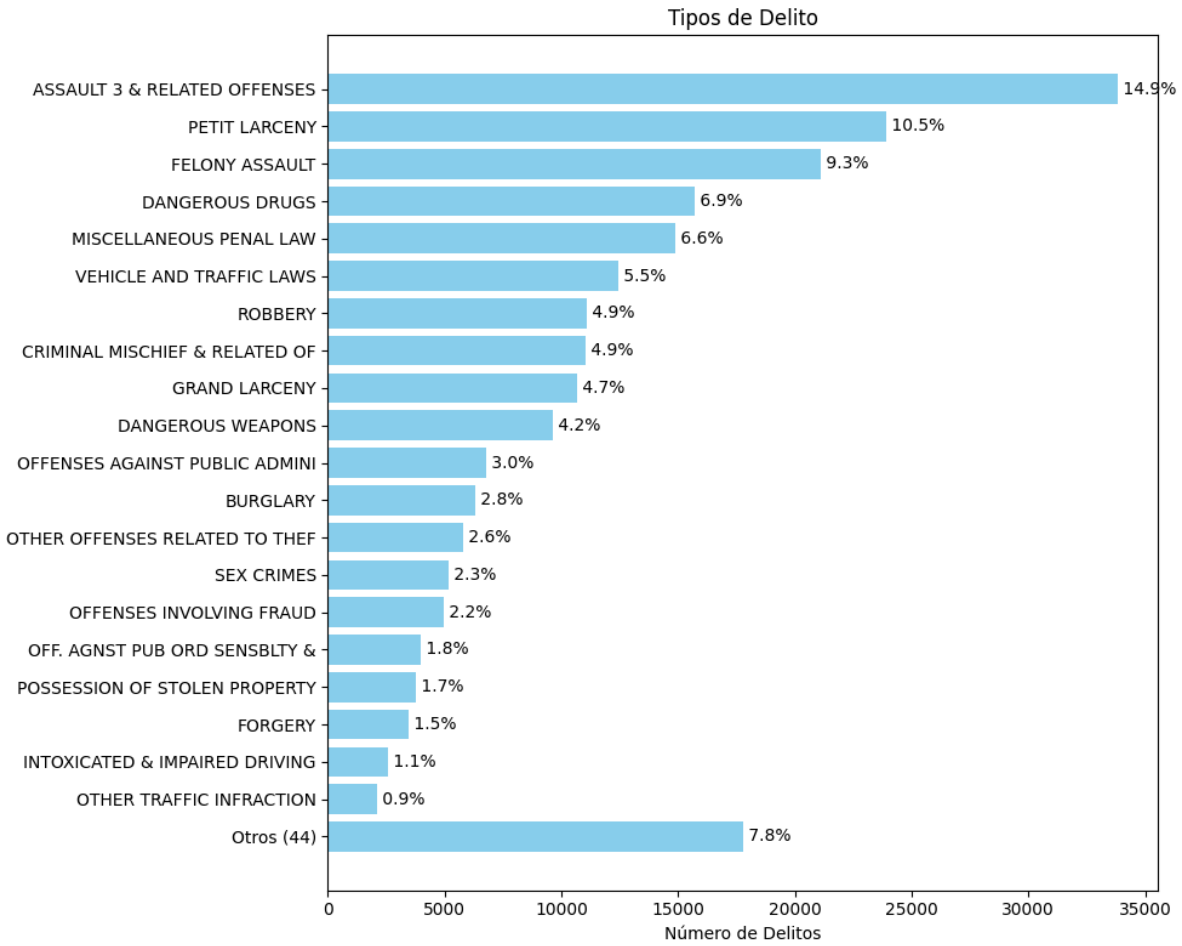
TIPO_DAÑO_PROPIEDAD_PÚBLICA	Tipo de propiedad pública dañada (por ejemplo, señal, cerca, poste de luz, etc.)	Cadena
FACTOR_CONTRIBUYENTE_1	Factores que contribuyen a la colisión para el vehículo designado	Cadena
FACTOR_CONTRIBUYENTE_2	Factores que contribuyen a la colisión para el vehículo designado	Cadena

Este DataSet cuenta con 10 columnas más que el anterior y con una cantidad de registros 40 veces mayor que el DataSet de arrestos. Sin embargo, estas columnas demás tienden a ser cadenas de caracteres/ párrafos que describen la colisión. Por lo cuál, puede que no sean del todo utilizadas debido a que no se va a hacer un Procesamiento de Lenguaje Natural

IV. Análisis Exploratorio de Datos

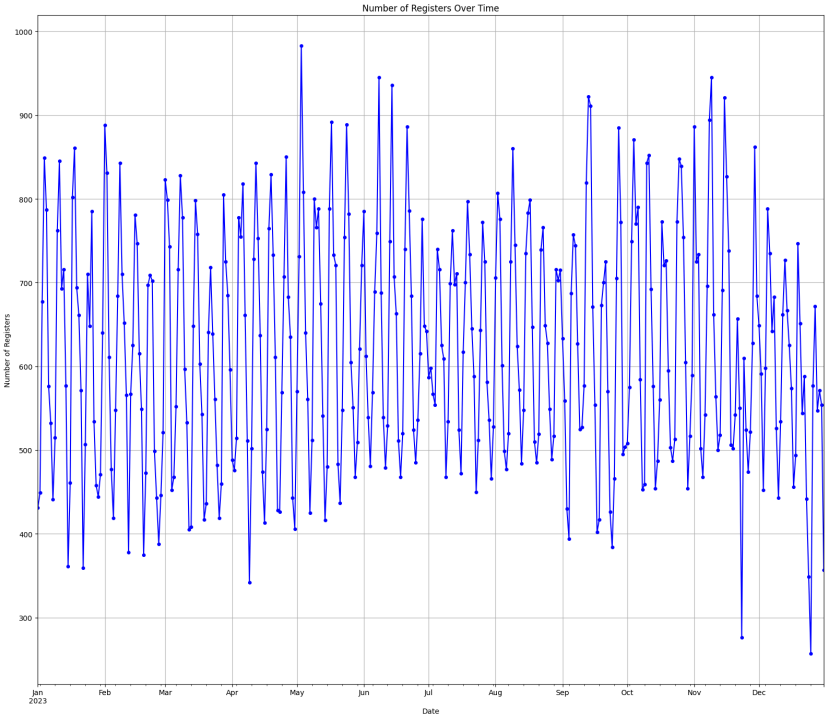
A. DataSet Arrestos

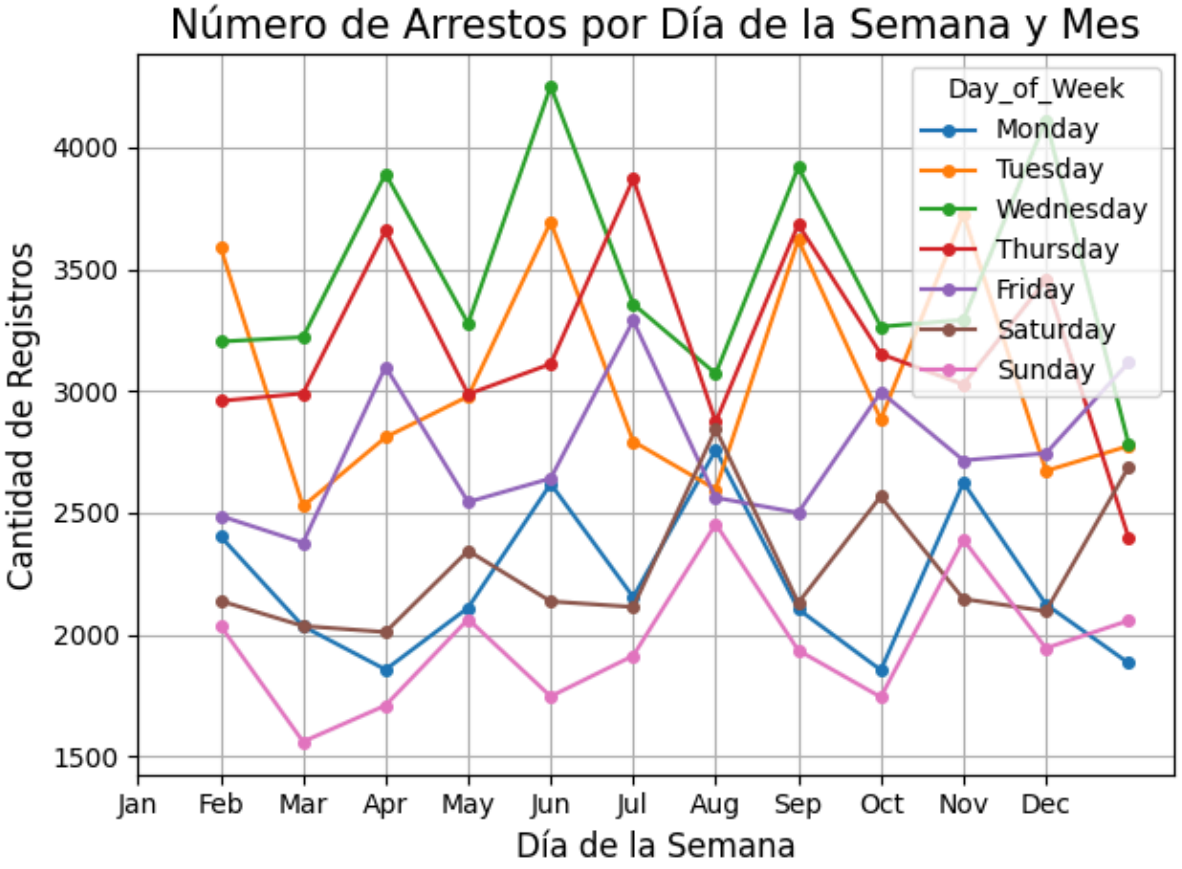
Para este primer DataSet se pensó primero en entender cuáles eran los delitos que tenían mayor cantidad de registros:



Aquí se encontró que los delitos están bastante distribuidos en el sentido de que no hay un delito que tenga más de un cuarto de los registros actuales. Ahora, esto no significa que no hayan delitos que tuvieron mayor incidencia que otro.

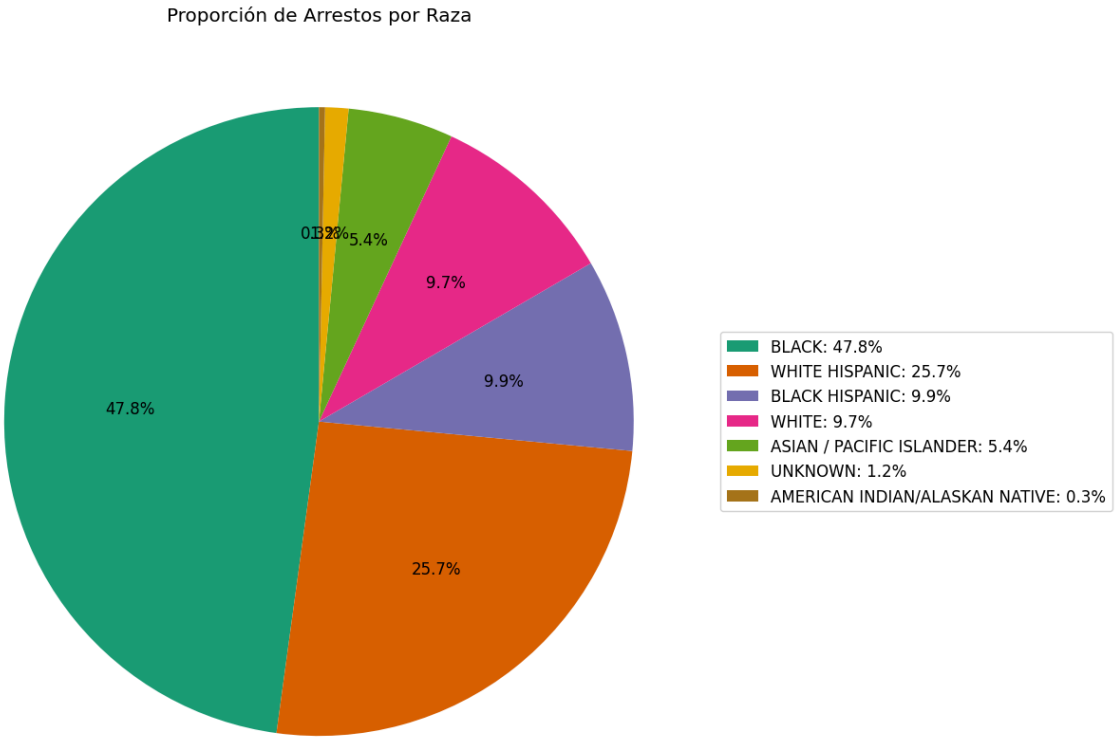
También se buscó ver alguna tendencia en las horas en las que ocurrían más arrestos en el día:





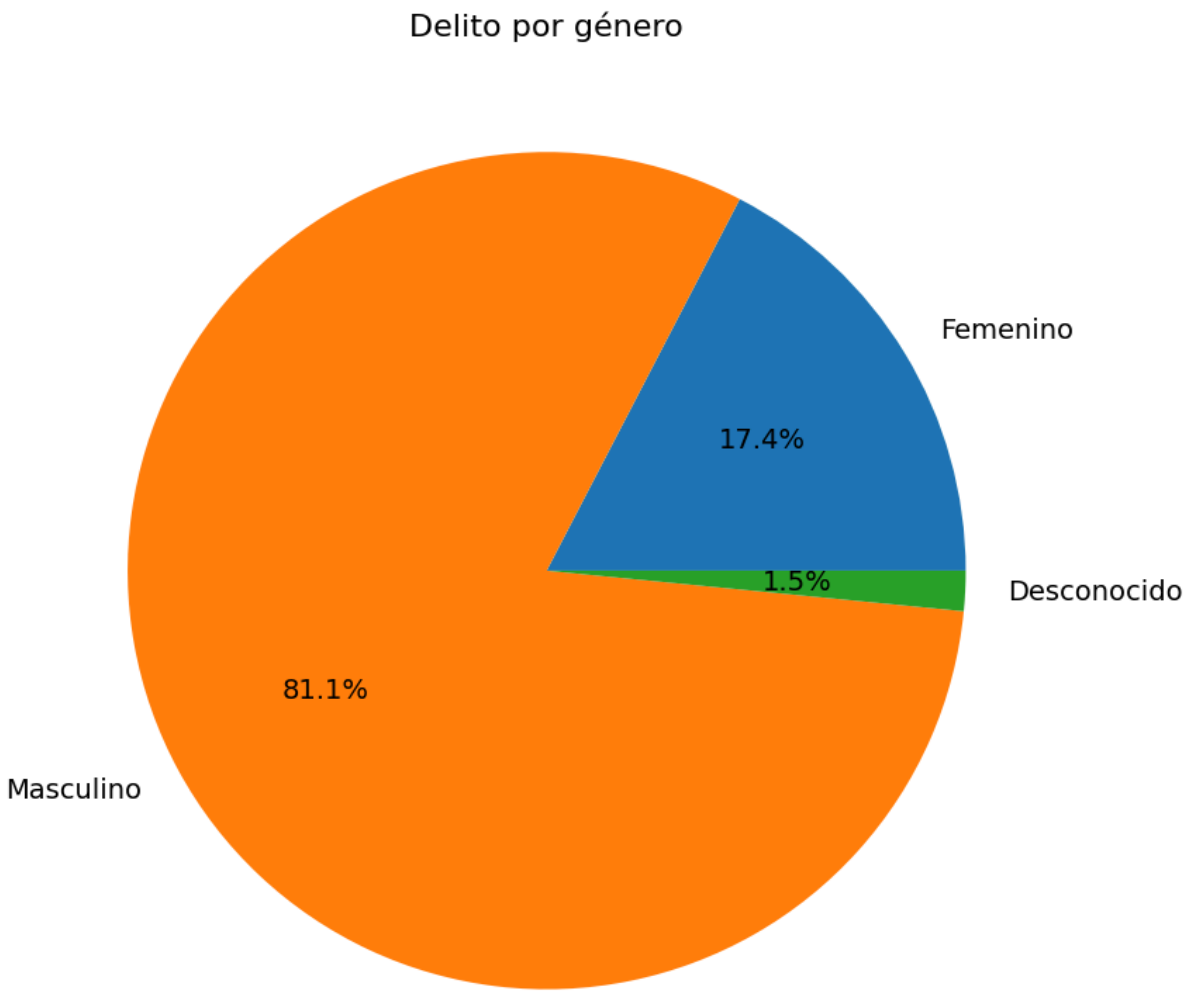
Sin embargo, como se puede observar es bastante uniforme la distribución cómo para poder sacar alguna conclusión respecto a cuál(es) es(son) las horas en las que se cometen más delitos. Sin embargo, no se descarta el hecho de que se podría sectorizar esta estadística por delito y que se pudieran hallar *insights* respecto a al incidente de cada delito por separado.

Después se hizo uso de la columna de raza para ver si había alguna(s) raza que tuvieran un mayor número de registros y se encontró lo siguiente:



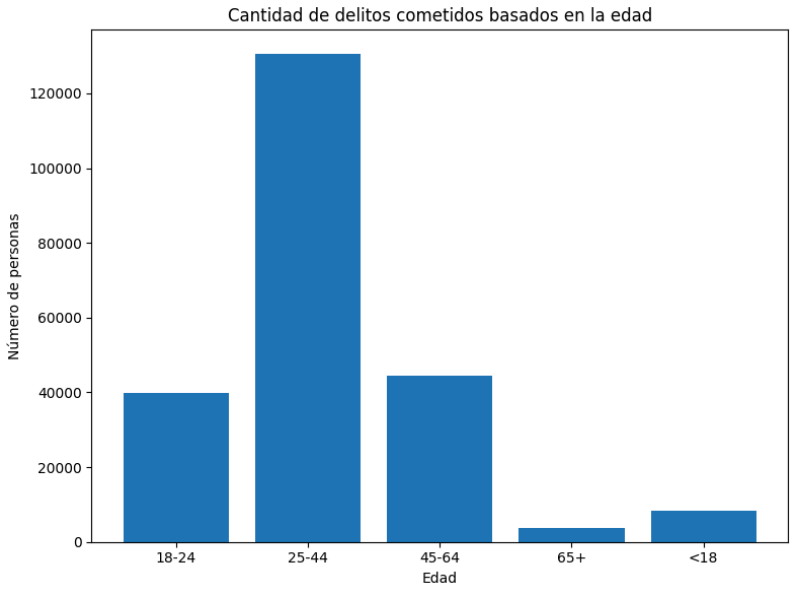
Más del 80% de los arrestos fue a personas Negras, Hispanas de color blanco o hispanas de color negro. En este sentido, serían importante poder revisar los datos poblaciones de New York para ver si en realidad estas razas constituyen este número de población de NYC, o si por el contrario, son un minoría en la población total que son la mayoría en los arrestos, lo cuál podría ser preocupante.

También se utilizó una diagrama de Pie para poder ver sí había algún sexo que tuviera una mayor porcentaje de registros y se enocontró lo siguiente:



Cómo se puede observar, los hombres tienen mayor incidencia que las mujeres por una mayoría aplastante (80% contra un 17%) respectivamente. Sin embargo, se podría hacer una revisión más profunda para entender qué tipo de delitos son los que más cometen los hombres y cuáles son los delitos que más cometen las mujeres, en caso de que los hombres cometieron más delitos de un tipo y las mujeres de otro.

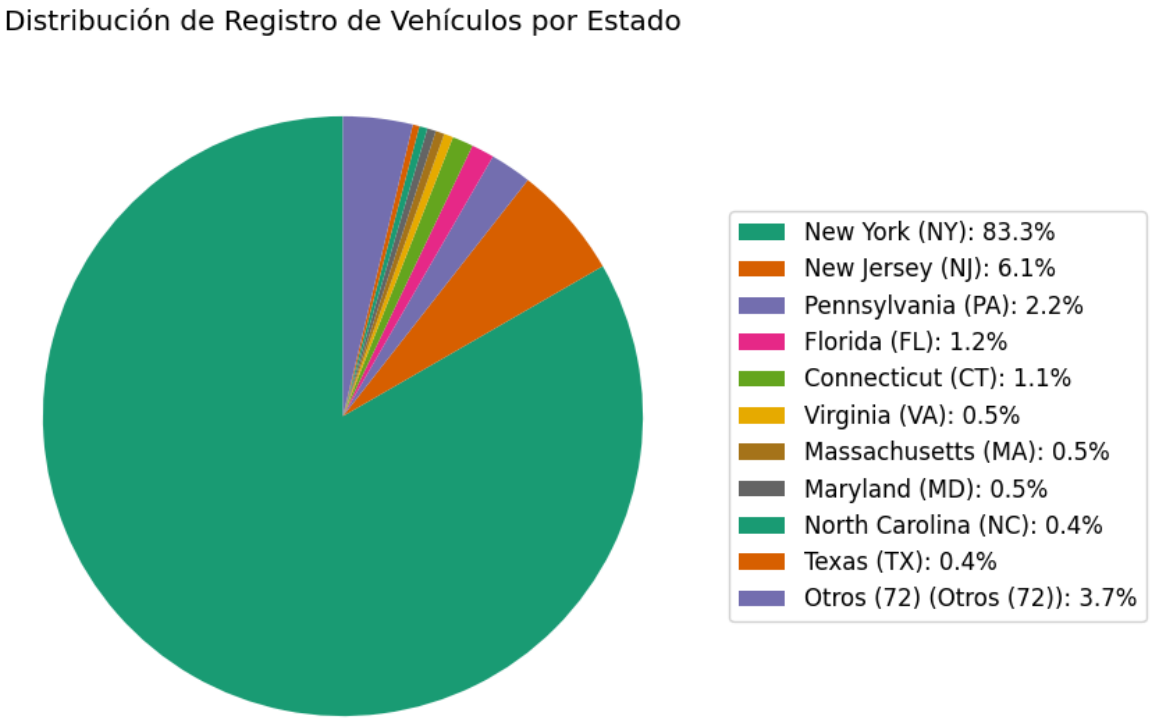
Por último, se hizo un histograma para poder entender si había un rango de edad en dónde se tuviera un mayor nivel de arrestos:



Cómo se puede observar, hay una mayor cantidad de personas entre 25-44 años que han sido arrestadas a comparación del rango etario de <18,18-24, 45-64 y 65+. Sin embargo, puede que ciertas personas en un rango de edad sean más propensas a cometer cierto delito.

B. DataSet Colisiones

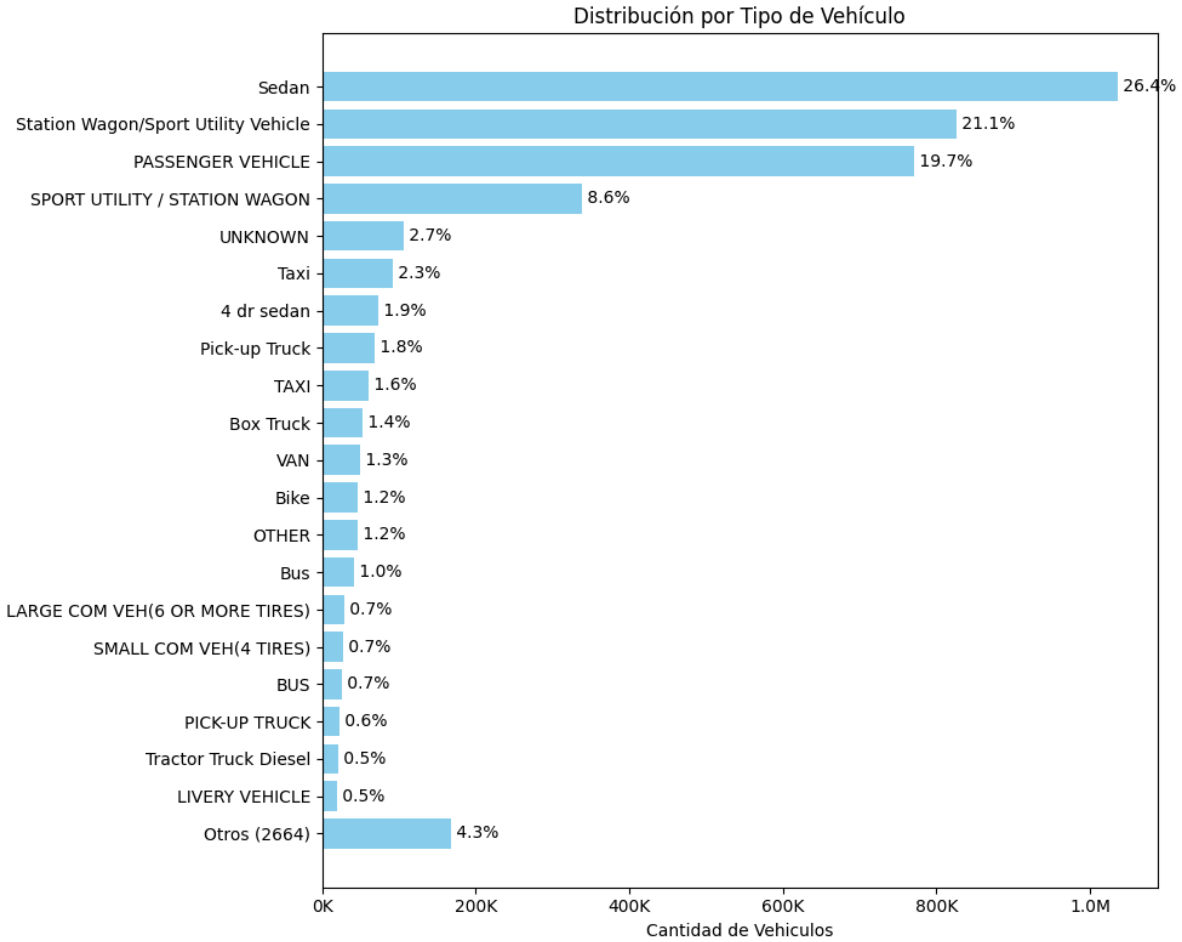
Para este primer DataSet se pensó primero en entender de dónde provenían este tipo de accidentes, es decir, si los accidentes viales eran producidos por personas que vivían dentro de New York o si estos eran producidos por personas que pasaban por New York para llegar a sus destino final.



Si bien en algunas partes del gráfico hay ciertos nombres sobrepuestos, estos nombres no tienen mayor relevancia debido al bajo porcentaje que se le asigna al estado de viene la placa del automóvil. Sin embargo, aquellos estados o lugares de dónde se tiene un mayor porcentaje son el propio estado de New York con el 83.3%, New Jersey con 6% y Pennsylvania con el 2.2%.

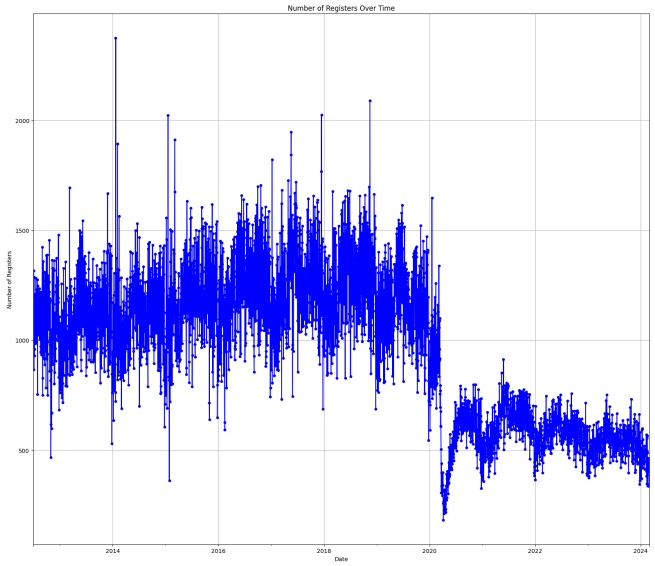
Lo que se puede observar es que los siniestro son causados por personas de la zona o individuos que viven en estados contiguos a New York, esto se puede observar en la siguiente imagen:

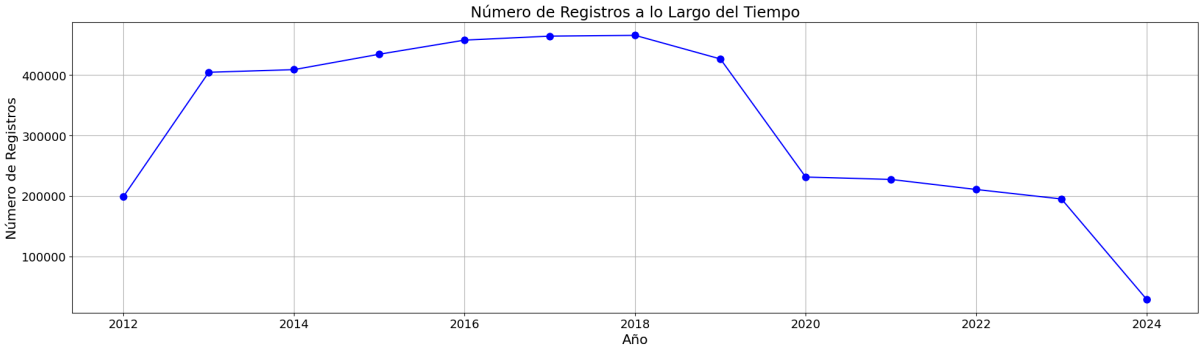
Otra visualización que se hizo fue poder hallar qué tipo de vehículos se accidentaron más, para lo cuál se encontró:



Para lo cuál se encontró que los tres vehículos que más se accidentaron fueron los carros SUV (camionetas deportivas), los sedanes y los vehículos de pasajeros diferentes a los buses, taxis, etc. Estos vehículos constituyen más del 83% de las colisiones.

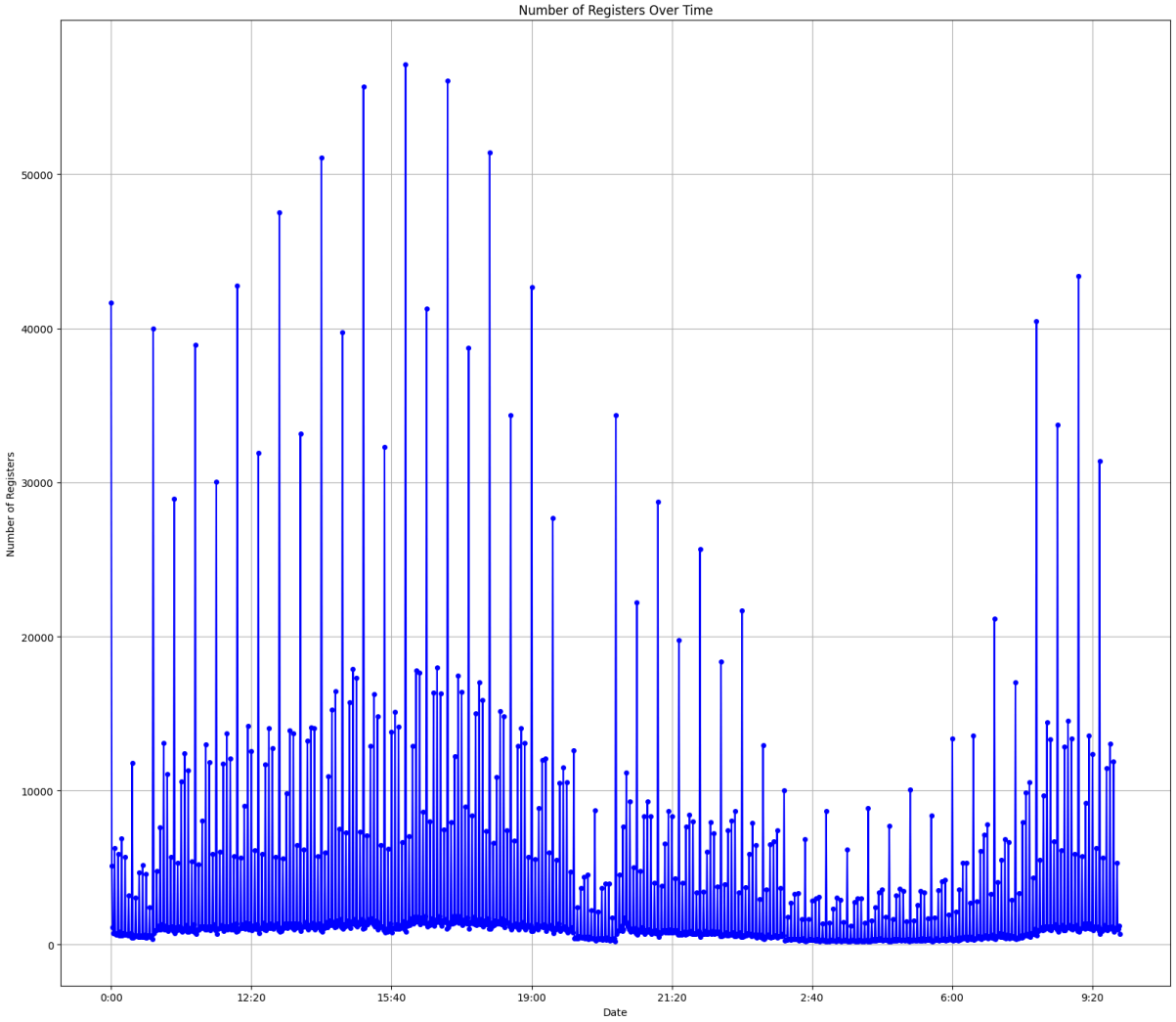
Otro visualización que se hizo fue el número de registros de colisiones sobre el tiempo y lo que se encontró fue lo siguiente:

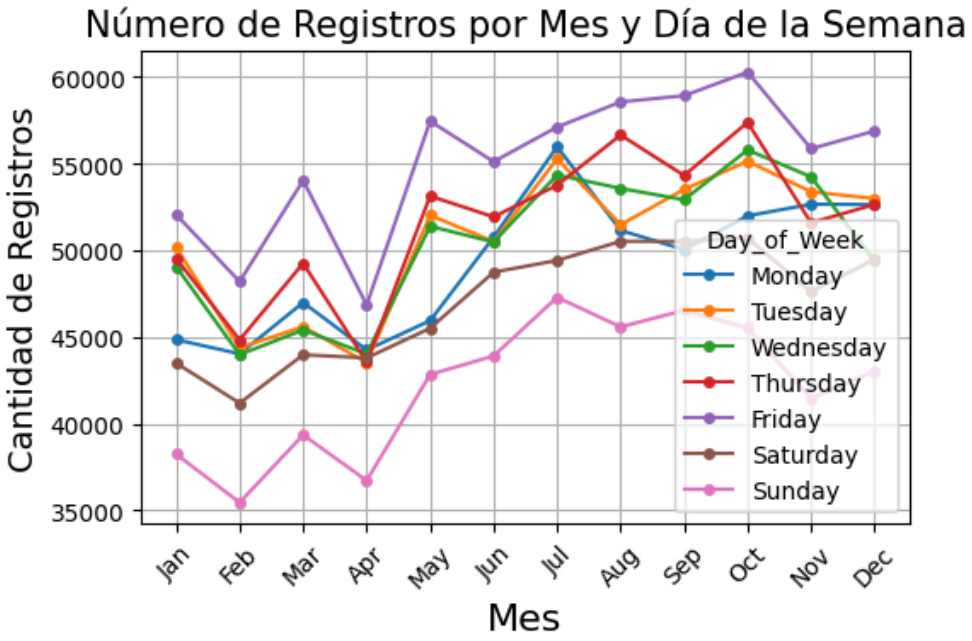




Lo primero que se encontró fue el hecho de que después de la pandemia de 2020 hubo un decremento sustancial y casi que permanente en el número de registros. También, si revisa la gráfica con detalle, se puede observar un comportamiento cíclico cada año, en dónde hay menor incidentes entre el inicio y el final del año, pero a mitad de año se tiene el pico de accidentalidad. Es decir, que se podría entrenar un modelo el cuál podría predecir el pico de accidente de un año dado.

Se hizo una gráfica en la cuál se buscó saber que horas del día eran las que tenía más registros para así saber cuál era la hora de mayor accidentalidad en el día:





Si se lograran agrupar los datos se podría evidenciar una tendencia más contundente respecto a los horarios en dónde hay mayores niveles de colisiones.

V. Reporte de calidad

Para los valores faltantes se encontró en el dataset de colisiones se halló lo siguiente:

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

Para los valores faltantes en el dataset de arrestos se encontró lo siguiente

Nombre Columna	Número de Nulos
ARREST_KEY	0
ARREST_DATE	0
PD_CD	2
PD_DESC	0
KY_CD	17
OFNS_DESC	0

LAW_CODE	0
LAW_CAT_CD	0
ARREST_BORO	1599
ARREST_PRECINT	0
JURISDICTION_CODE	0
AGE_GROUP	0
PERP_SEX	0
PERP_RACE	0
X_COORD_CD	0
Y_COORD_CD	0
Latitude	0
Longitude	0
New Georeferenced Column	0

En el DataSet de Colisiones se utilizaron las siguientes técnicas para rellenar los valores faltantes:

a. En la columna DRIVER_LICENSE_JURISDICTION se encontró que esta guarda una correlación STATE_REGISTRATION, por lo cuál, si esta tenía un valor nulo, se utilizaba el valor de STATE_REGISTRATION en esa fila.
b. Basado en la columna VEHICLE_TYPE se hizo un diccionario que almacenaba el valor que tenía esta columna para un valor en VEHICLE_MODEL, ya con esto, en caso de que se tuviera un valor de VEHICLE_TYPE y no se tuviera VEHICLE_MODEL, se procedía a reemplazar el valor faltante basado en el diccionario.
c. Debido a que las columnas de VEHICLE_DAMAGE_X contaban con categorías del tipo de daño hecho, se optó por convertir estas categorías en columnas binarias (en dónde 1 significa que si se sufrió ese daño y 0 que no) y eliminar las 4 de VEHICLE_DAMAGE, esto también se hizo porque se espera que en la próxima entrega se pueda utilizar una de las columnas de la categoría cómo la variable objetivo para luego entrenar y testear un modelo.

Para el dataset de Arrestos se utilizaron las siguiente técnicas para rellenar los valores faltantes:

a. Para la columna KY_CD se generó un diccionario de los valores equivalentes en PC_CD para luego, utilizar los valores que tiene PC_CD en KY_CD nulo, reemplazando los valores basado en el diccionario de equivalencias en valores.

VI. Planteamiento de preguntas de negocio

Para el planteamiento de las preguntas de negocio se pensaron en las siguientes:

- DataSet Colisiones:
    - ¿Cuaĺes son las horas o intervalos de tiempo en dónde se presenta mayor accidentalidad (y en qué zonas)?
    - ¿ Cuáles son el top 3 de vehículos más propensos a generar accidentes?
    - ¿ Hay algún perfil (conjunto de características) en específico que pueda generar mayores niveles de accidentalidad?
    - ¿ Se podría predecir el tipo de daño o la probabilidad de algún tipo de daño basado en algún perfil?

- DataSet Arrestos:
    - ¿ Se podría, basado en un perfil y zona, predecir el crimen ocurrido?
    - ¿ Hay algún perfil específico que tenga mayor probabilidad de cometer
        - Robbery.

- Grand Larceny.
- Hate Crimes.
- Felony Assault.
- Grand Larceny Auto.

- ¿ En qué latitudes se cometen qué cantidad de crímenes con más frecuencia?
- ¿ Qué razas de qué género cometen los crímenes antes mencionados?

VII. MODELADO

Para el modelado se utilizaron dos tipos de modelos: Random Forest y Kmeans.

En el caso de Random Forest se utilizó este modelo para predecir ciertos crímenes en específico y también en qué casos se podría decir que un crimen fue cometido por una persona afrodescendiente y que crimen no.

Para el primer modelo de Random Forest se utilizaron las variables del Borough, la edad, la raza y el sexo para poder predecir si dicho perfil era de hurto o no.

```
Accuracy: 0.6774415906005872
```

Feature	Importance
ARREST_BORO_index	0.0
AGE_GROUP_index	0.705927326537091
PERP_RACE_index	0.10600987650473018
PERP_SEX_index	0.016409353539047225
ARREST_PRECINCT_i...	0.17165344341913172

Como se puede observar este modelo tuvo una métrica del 0.67 de precisión y le dio más importancia a la edad junto con la raza y el sexo. En este caso en específico el borough no tuvo mayor incidencia en sí una persona cometía un hurto o no.

Después se trató de predecir si dadas ciertas características se cometía el delito de *Grand Larceny*, las columnas que se usaron también fueron el *borough,* la edad, la raza, el sexo y el *precinct*.

```
Accuracy: 0.63680643047272
```

Feature	Importance
ARREST_BORO_index	0.07197880342248002
AGE_GROUP_index	0.0698596402142934
PERP_RACE_index	0.016618526060182204
PERP_SEX_index	0.04019815108182278
ARREST_PRECINCT_i...	0.8013448792212217

Como se puede observar el modelo tomó cómo variable principal el *precinct* seguida del *borough*, algo que es bastante diferente con respecto al caso de *Robbery* debido a que en ese caso la edad y la raza tienen mayor importancia mientras que acá pasan a un segundo plano. Sin embargo, la precisión de este modelo es menor al anterior, por lo cuál no se podría decir que al tomar estas variables se pueden tener mejores resultados.

Por último, se intentó predecir que dadas ciertas características de una arresto, poder predecir si este fue hecho a una personas de raza negra o no. Esto, con el fin de darle herramientas a estas personas para poder utilizar la tecnología y el modelado para presentar evidencia contundente de último recurso en caso en dónde ya vayan a condenar a la persona a pesar de que sea inocente pero que no cuente con suficientes evidencias para probarlo. Para este modelo se utilizaron las siguientes métricas:

```
Accuracy: 0.641106264254556
Precision: 0.6406926300932817
Recall: 0.641106264254556
F1 Score: 0.6405882332556421
AUC: 0.7041555013829053
```

Se utilizaron las columnas de *Borough,* del tipo de delito/ofensa y el *arrest precinct* el cuál es el cuadrante, es decir, una parte o ubicación más específica que el borough. Para esto se obtuvieron lo siguientes importances:

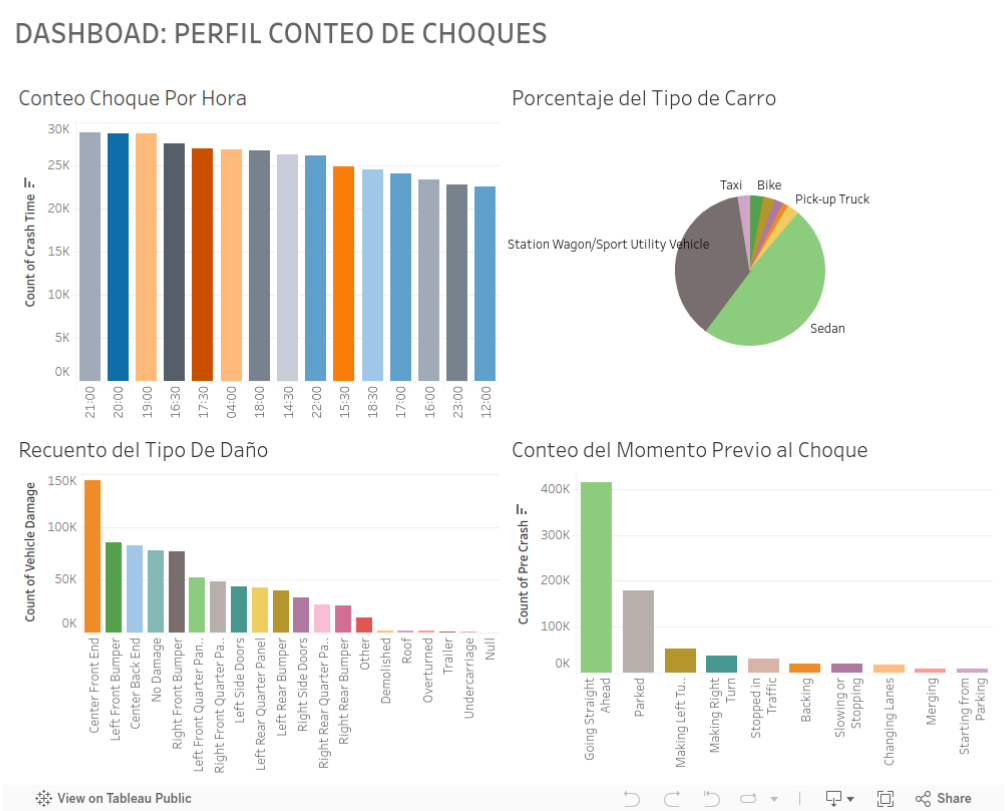
```
+--------------------+--------------------+
|             Feature|          Importance|
+--------------------+--------------------+
|   ARREST_BORO_index| 0.18409651741806976|
|     OFNS_DESC_index|0.010737294455013716|
|ARREST_PRECINCT_i...|  0.8051661881269165|
+--------------------+--------------------+
```

En este caso el precinct tuvo una mayor importancia para el modelo, sin embargo, como se puede ver en las métricas del modelo, este tiene unos resultados muy bajos como para ser utilizado en escenarios tan comprometedores cómo lo es la libertad que puede tener una persona.

VIII. VISUALIZACIONES Y DASHBOARDS

Los *dashboards* que se van a presentar a continuación fueron diseñados de tal forma que se puede experimentar cómo afecta un parámetro el resto de los datos, en este sentido cada *dashboard* tiene una visualización que afecta al resto. En los tres tableros esta gráfica es la que se encuentra en la esquina superior izquierda, ahora, se van a presentar algunos screenshots de estos tableros, pero la idea de estos tableros es poder facilitar al gerente/*manager* el poder estudiar los datos de una manera visualizar, interactiva y agradable a la vista.

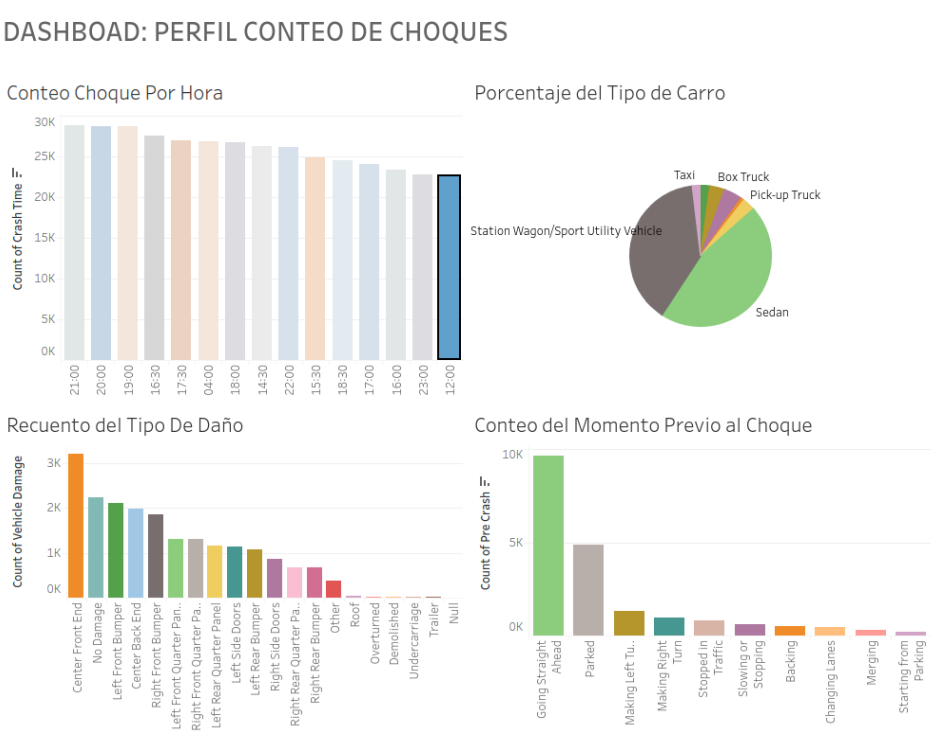
El primer *dashboard* que se quiere presentar es el del perfil del choque basado en el conteo por hora:



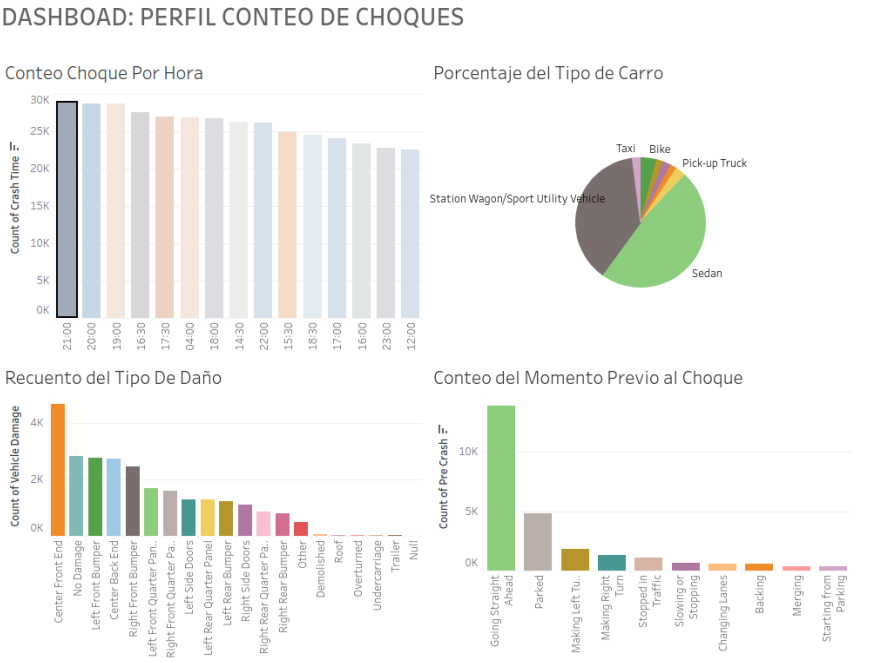
Este dashboard tiene como "parámetro" el conteo de choques por hora, está organizado de manera descendente y contiene un top 13 de las horas en dónde hay mayor cantidad de choques. Dependiendo de la

barra que se selecciona se ve afectada la gráfica que tiene un diagrama de torta que contiene el tipo de carro/vehículo, después está la gráfica que contiene el conteo del tipo de daño ocasionado (el top 15 basado en el parámetro) y por último el conteo de la situación en la que se encontraba el conductor antes de que ocurriera el choque.

Una de las gráficas "parametrizadas más interesantes la del conteo de choques a las 12 m:

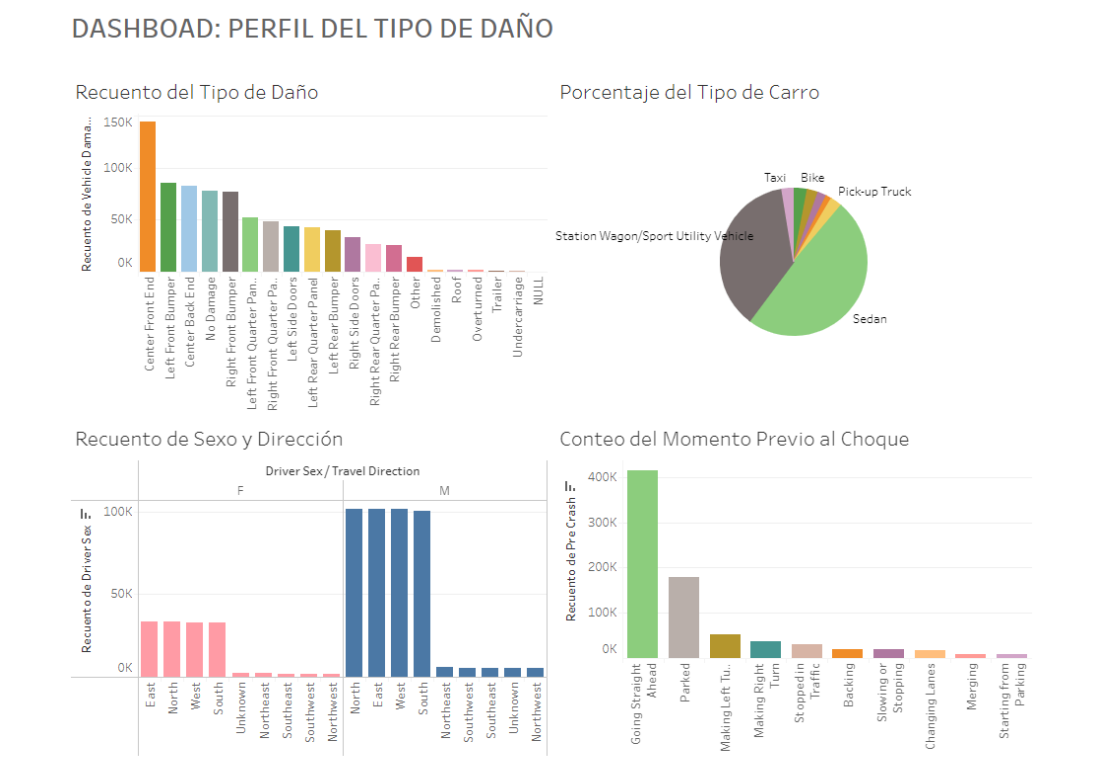


Cómo se puede ver a esta hora aumenta el recuento de *box trucks* por choque junto con el hecho de que en esta hora *no damage* ocupa la segunda posición en en el tipo de daño así cómo el hecho de que se ve que en este caso del top 4 del conteo del momento previo al choque se mantiene intacto.

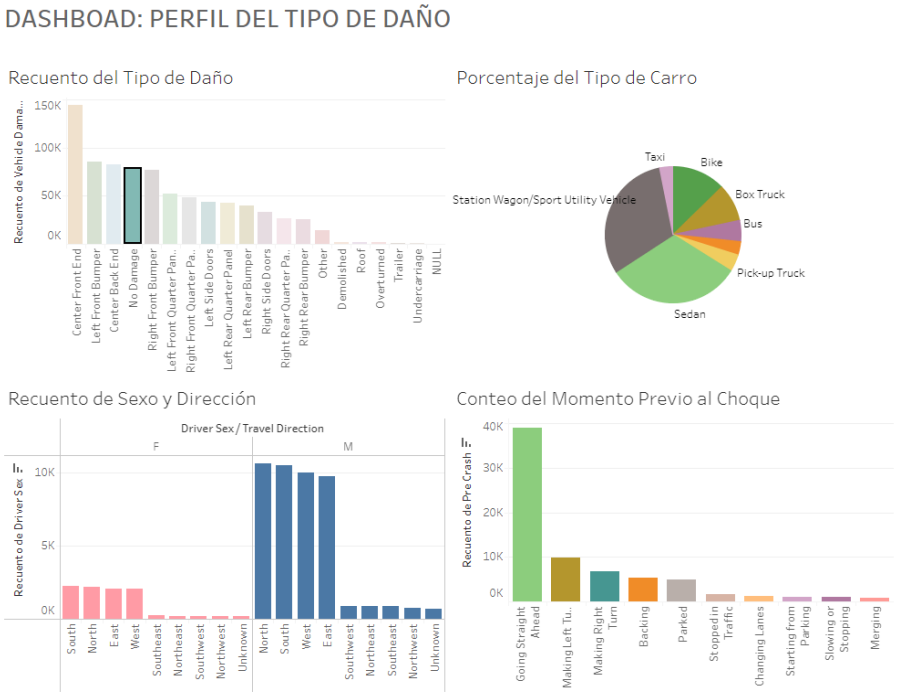


En este tablero de los choques a las 9 pm (21:00) se ve que hay una tendencia al conteo general muy similar, sin embargo, se puede ver que *No Damage* también se encuentra en la segunda casilla del recuento del tipo de daño dentro de estos choques además se ve que en este caso el Box Truck tiene un menor recuento que el gráfico de las 12 m.

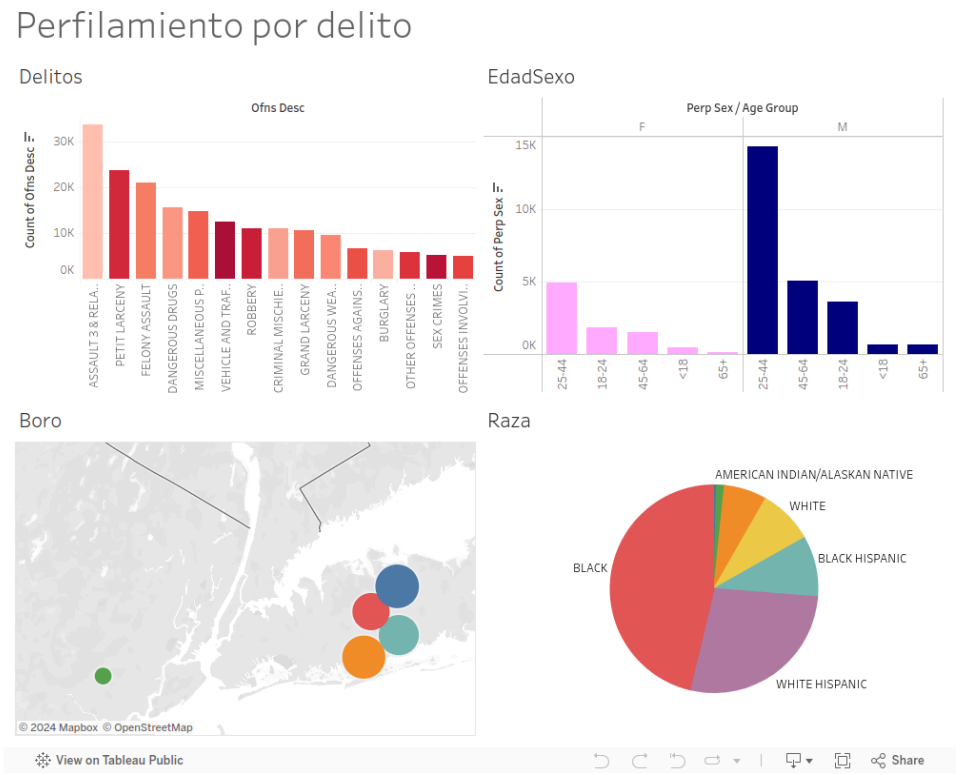
El segundo tablero que se tiene consiste en tener un perfil basado en la parte del vehículo en dónde se tuvo el daño, es decir, el tipo de daño. Basado en esto se quiere saber el tipo de vehículo que lo sufrió el sexo y dirección hacia la que se dirigía el conductor y por último un histograma que logre mostrar un conteo del momento previo al choque:



Uno de los casos en los que más resaltan es el de la categoría *No Damage* debido a que en esta categoría de daño se presenta un fenómeno en dónde se distribuye en otros vehículos la cantidad de choques. En este caso vemos como las bicicletas, los buses y los *bus trucks* empiezan a tener mayor protagonismo en este categoría. Además, en este caso puntual, el hecho de estar parqueado tuvo poco efecto en este tipo de colisiones, en cambio, el hecho de ir en reversa o como dice en la gráfica *backing* si tuvo incidencia en este tipo de accidentes.

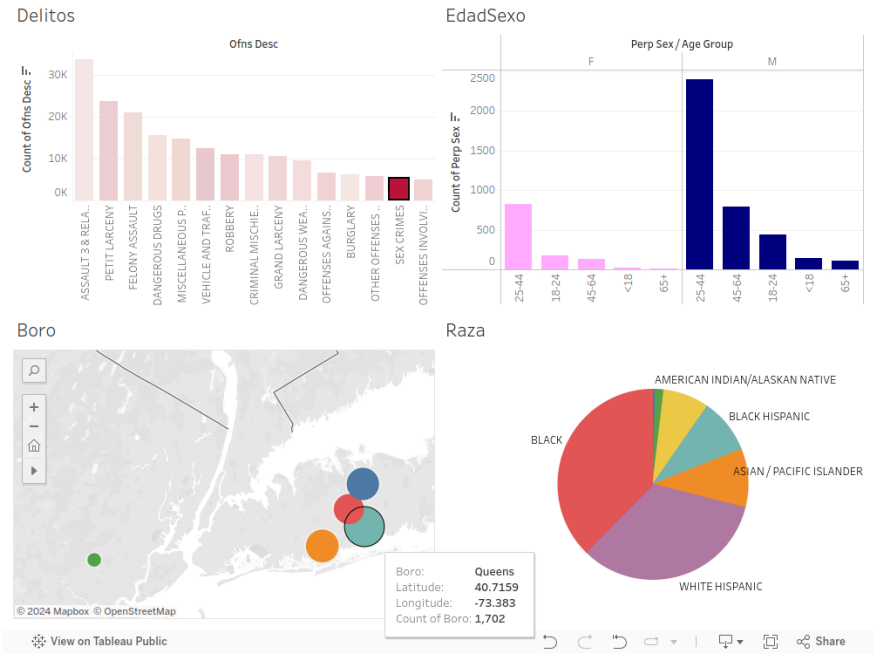


El siguiente tablero es que contiene todo lo relacionado con delitos:



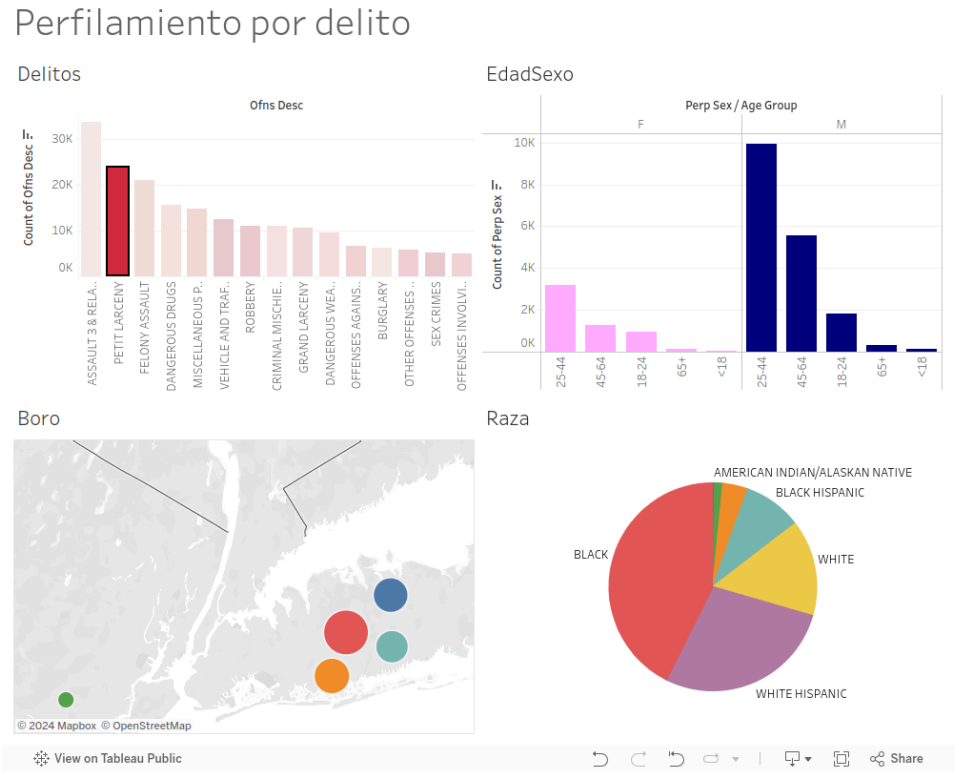
Este dashboard se compone de 4 gráficas: la primera gráfica es un histograma el cuál tiene ordenados de manera ascendente el número de veces que se repite un delito específico, en la gráfica en la esquina superior derecha se encuentra una gráfica que muestra el conteo de estos delitos divididos por edad y por sexo. En la esquina inferior izquierda vemos un mapa el cuál muestra los 5 *boroughs* de NY, en esta gráfica el conteo de estos delitos se refiere al tamaño del círculo. Por último se tiene un diagrama de torta que contiene como se dividen por raza estos delitos.

Uno de los crímenes que tiene comportamiento atípico son los crímenes relacionados al sexo:



En este tipo de crímenes las dos tendencias más claras son el hecho de que hay un mayor recuento de estos crímenes en Queens y también el hecho de que en estos crímenes la tendencia general de la raza no

se cumple, de hecho, resalta el hecho de que la tercera raza que más se involucra en estos crímenes son las personas que son asiático-americanas.



La siguiente categoría son los hurtos menores, en esta categoría resalta el hecho de que la mayor cantidad de estos hurtos ocurren en Manhattan y que la tercera raza la cuál comete este delito (diferente a la tendencia general) es la de color blanco.

IX. LIMPIEZA Y TRANSFORMACIÓN INICIAL

Para la limpieza de cada DataSet se tomaron enfoques diferentes pero si hubieron eliminaciones de columnas que no se consideraron necesarias para el análisis.

DataSet de Arrestos:

- Se eliminó la columna ‘KY_CD’ debido a que la información de esta era almacenada por la columnas ‘PD_CD’ de una forma diferente, pero que hacía alusión a lo mismo.
- Se eliminó la columna ‘ARREST_KEY’ debido a que solo ayuda a diferenciar un registro de otro, y cómo se halló que no hay registros repetidos, los valores de esta columna ya no pueden aportar mayor información al análisis.
- Se eliminó la columna ‘JURISDICTION_CODE’, esta columna se eliminó porque no se pudo interpretar el significado que esta tenía, no había información sobre esta en el diccionario y tampoco era una columna clave para el análisis.
- Se eliminó la columna ‘LAW_CODE’ debido a que a pesar de poder tener datos interesantes respecto a la sentencia que se le da a la persona arrestada, esta variable no se le vio mayor uso para predecir o para analizar.
- Se eliminó la columna 'New Georeferenced Column' debido a que el valor de esta columna ya se podía obtener con Latitude y Longitude, por lo que se optó por eliminarlo por lo que no se utilizaría en el análisis sino los valores de Latitude y Longitude por separado.
- Se eliminó la variables LAW_CAT_CD debido a que no se le vio un uso para poder ser graficada o usada en un modelo predictivo.

En el DataSet de Colisiones:

- Se eliminó la columna ‘UNIQUE_ID’ debido a que solo ayuda a diferenciar un registro de otro, y cómo se halló que no hay registros repetidos, los valores de esta columna ya no pueden aportar mayor información al análisis.
- Se separó la columna ‘VEHICLE_MAKE’ en MODEL y MAKE dónde una almacena la marca del carro y la otra almacena el tipo de carro. Posteriormente se eliminó la columna inicial.

- Se eliminó la columna ‘MODEL’ debido a que la información de esta columna era la misma y más completa (tenía una menor cantidad de nulos) en la columna de ‘VEHICLE_TYPE’.
- Se eliminó la columna ‘VEHICLE_ID’ debido a que no se podía utilizar en análisis ni el la predicción.
- Se eliminaron las columnas ‘PUBLIC_PROPERTY_DAMAGE’ y ‘PUBLIC_PROPERTY_DAMAGE_TYPE’ debido a que contaban con muchos valores nulos y porque no eran variables objetivo.
- Se eliminó la columnas ‘VEHICLE_MODEL’ debido a contenía muchos valores nulos los cuáles no tenían la posibilidad de ser reemplazados o imputados por algún valor coherente.
- Las horas de registro en la columna ‘CRASH_TIME’ se cambiaron a un formato en dónde dependiendo del valor de la hora esta pertenece a las horas :30 o a las horas :00, es decir, que si antes había un registro que contaba con el valor de ‘10:28’ ahora, este será de ‘10:30’. En este sentido, se puede garantizar que se va a tener una gráfica más clara al momento de saber cuáles son las horas/ intervalos de tiempo, en dónde ocurren más accidentes.
- Se estandarizó la variable CRASH_TIME a formato timestam, es decir, HH:mm.
- Se estandarizó la variables CRASH_DATE a formato date, es decir, dd/mm/yyyy.
- Se obtuvo un sub conjunto de datos el cuál contenía solo la información de 4 años hacia atrás para poder ver cómo ha sido el cambio o los efectos del actual alcalde de Nueva York en las colisiones

X. Conclusiones

La primera conclusión que se puede obtener de este ejercicio es el hecho de que ciudades como *New York* que tienen ingresos y economías de gran magnitud, no están exentas de la inseguridad y de los siniestro viales. A pesar de que las métricas de seguridad según el departamento de policía de la ciudad, van en mejoría, no todos estos indicadores muestran todos los problemas de seguridad que la ciudad tiene actualmente estén solucionados.

Respecto a los delitos y las conclusiones que se obtuvieron del conjunto de datos de arrestos se puede empezar concluyendo que los tres delitos que más ocurren en esta ciudad son los asaltos de tipo 3 y ofensas relacionadas, el hurto menor y los delitos graves de asalto. Además, los tres días en dónde se prestan mayor cantidad de crímenes son los miércoles, Jueves y Martes, mientras que los dos días dónde se presentan menos crímenes son los lunes y los Domingos.

Si bien la raza y el género pueden incidir de cierta manera respecto a si una persona comete o no un crimen, no se puede concluir de manera apresurada que las personas de raza negra y que los hombres son los mayores responsables de los crímenes en la ciudad, en este sentido se debería poder contar un conjunto de datos de enjuiciamiento en dónde se puede verificar que proporción de hombres y personas de raza negra son declaradas culpables despueś de cometer el delito. Sin embargo, una conclusión demográfica concluyente es el hecho de que a edad de las personas que cometen crímenes dentro de esta ciudad oscila entre los 25 y 44 años de edad, es decir, que no son las personas más jóvenes ni las personas más viejas las que cometen estos delitos, sino las personas que se encuentran en la “mejor condición física posible” para poder cometer estos delitos.

En el caso del conjunto de datos de Colisiones pudimos concluir que la mayoría de los accidentes dentro de esta ciudad son causados por habitantes de esta o de alrededores, más los incidentes no tienen ningúna relación con personas provenientes de otros estados de la nación. También se puede afirmar el hecho de que los dos vehículos que, por lo general, generan más accidentes dentro de la ciudad son los carros sedanes y las camionetas SUV, sin embargo, hay casos y excepciones en dónde esto no se cumple, uno de estos casos es en las colisiones en dónde no hay daño, en este caso se tiende a ver las bicicletas, los *box trucks,* entre otros vehículos, tienden a llevarse parte del “protagonismo” en esta clase de incidentes.

Otra conclusión que se puede obtener del conjunto de datos de colisiones es el hecho de que la cantidad de estos a lo largo del tiempo muestra que después de la pandemia hubo una reducción bastante abrupta. Si bien no se saben las razones, si podría ser una muy buena pregunta de investigación saber el porqué de esta reducción. Junto a esto, se puede ver que los días en dónde ocurren más de estos incidentes son los viernes y los jueves, si bien se podrían sacar conclusiones apresuradas respecto a que la vida nocturna dentro de la ciudad genera una mayor cantidad de colisiones, es importante poder investigar más a fondo porqué esta tendencia ocurre en estos días.

References

[1] https://www.nyc.gov/office-of-the-mayor/
[2] https://es.statista.com/estadisticas/598684/producto-interior-bruto-pib-real-per-capita-en-los-ee-uu-por-estado-en/
[3] https://www.larepublica.co/globoeconomia/las-10-ciudades-del-mundo-que-por-su-economia-son-similares-a-un-estado-2556746

[4] https://www.fox5ny.com/news/nyc-crime-rate-2023-statistics
[5] https://www.nyc.gov/site/nypd/news/p00099/nypd-january-2024-citywide-crime-statistics
[6] https://www.mapsofworld.com/usa/east-coast-usa.html
[7] https://jknylaw.com/new-york-car-accident-lawyer/statistics/#:~:text=In%202020%2C%20the%20year%20of,were%20100%2C508%20accidents%20in%20NYC.
[8]