Proyecto Procesamiento de Datos

Nicolas Samuel Martin

Juan Diego Gonzalez

Juan David López Becerra

Pontificia Universidad Javeriana

Procesamiento de Datos a Gran Escala

22/05/2024

PRIMERA ENTREGA

INTRODUCCIÓN

"Optimización de la Seguridad Vial y Eficiencia de Conducción en Entornos Urbanos: Un Análisis de Datos de Colisiones para la Innovación de Tecnologías de Vehículos Autónomos de Tesla"

En la era de la movilidad inteligente y sostenible, Tesla se ha posicionado como un líder indiscutible en la revolución de los vehículos eléctricos y la conducción autónoma. La compañía se esfuerza constantemente por mejorar la seguridad, la eficiencia y la experiencia del usuario en sus vehículos. En este contexto, la ciudad de Nueva York, con su densa red de tráfico y alta incidencia de accidentes viales, ofrece un laboratorio urbano único para el desarrollo y la prueba de tecnologías avanzadas de conducción.

Reconociendo la complejidad de este desafío, Tesla ha iniciado una colaboración con (Nuestra consultora) para llevar a cabo un análisis exhaustivo de los datos de colisiones de vehículos en Nueva York. Este proyecto tiene como objetivo aprovechar los insights derivados de los datos para informar y optimizar las tecnologías de conducción autónoma y asistencia al conductor de Tesla, así como para identificar rutas de prueba óptimas que aseguren una navegación segura y eficiente en entornos urbanos.

La investigación se centrará en el análisis del conjunto de datos "Motor Vehicle Collisions - Vehicles" proporcionado por el Departamento de Policía de Nueva York (NYPD), que abarca detalles sobre cada vehículo involucrado en colisiones desde abril de 2016. A través de un enfoque analítico riguroso, este estudio buscará identificar patrones, factores contribuyentes y tendencias en los accidentes viales, con el fin de generar recomendaciones concretas para la mejora continua de las soluciones de movilidad de Tesla.

ENTENDIMIENTO DEL NEGOCIO

Tesla, reconocida por su innovación en vehículos eléctricos y tecnologías de conducción autónoma, busca continuamente mejorar la seguridad y eficiencia de sus productos. La ciudad de Nueva York, con su compleja red de tráfico y elevada incidencia de accidentes viales, presenta una oportunidad única para Tesla. Analizando los datos de colisiones de vehículos en esta metrópoli, Tesla puede identificar patrones y factores de riesgo específicos que informen el desarrollo y ajuste de sus sistemas de conducción autónoma y asistencia al

conductor. Además, este análisis puede contribuir a la planificación de rutas de prueba para sus vehículos, asegurando una navegación más segura y eficiente en entornos urbanos densos.

Para este proyecto, Tesla ha seleccionado el conjunto de datos "Motor Vehicle Collisions - Vehicles" proporcionado por el Departamento de Policía de Nueva York (NYPD). Este conjunto de datos es invaluable para comprender la dinámica de los accidentes viales en la ciudad y permite a Tesla adaptar sus tecnologías de manera más efectiva. La información detallada sobre cada colisión, incluyendo el tipo de vehículo, la dirección de viaje y los factores contribuyentes, es esencial para afinar los algoritmos de conducción autónoma y diseñar estrategias de prevención de accidentes específicas para entornos urbanos.

SELECCIÓN DE DATOS A UTILIZAR

Para este proyecto, Tesla ha seleccionado el conjunto de datos "Motor Vehicle Collisions - Vehicles" proporcionado por el Departamento de Policía de Nueva York (NYPD). Este conjunto de datos es invaluable para comprender la dinámica de los accidentes viales en la ciudad y permite a Tesla adaptar sus tecnologías de manera más efectiva. La información detallada sobre cada colisión, incluyendo el tipo de vehículo, la dirección de viaje y los factores contribuyentes, es esencial para afinar los algoritmos de conducción autónoma y diseñar estrategias de prevención de accidentes específicas para entornos urbanos.

Exploración de los Datos

Tablas:

Tabla de tipos de datos de los atributos

#	Column	Dtype
0	UNIQUE_ID	int64
1	COLLISION_ID	int64
2	CRASH_DATE	object
3	CRASH_TIME	object
4	VEHICLE_ID	object
5	STATE_REGISTRATION	object
6	VEHICLE_TYPE	object
7	VEHICLE_MAKE	object
8	VEHICLE_MODEL	object
9	VEHICLE_YEAR	float64
10	TRAVEL_DIRECTION	object
11	VEHICLE_OCCUPANTS	float64
12	DRIVER_SEX	object
13	DRIVER_LICENSE_STATUS	object
14	DRIVER_LICENSE_JURISDICTION	object
15	PRE_CRASH	object
16	POINT_OF_IMPACT	object
17	VEHICLE_DAMAGE	object
18	VEHICLE_DAMAGE_1	object
19	VEHICLE_DAMAGE_2	object
20	VEHICLE_DAMAGE_3	object
21	PUBLIC_PROPERTY_DAMAGE	object
22	PUBLIC_PROPERTY_DAMAGE_TYPE	object
23	CONTRIBUTING_FACTOR_1	object
24	CONTRIBUTING FACTOR 2	object

Tabla 1

Tabla estadística por tipo de vehículo

VEHICLE_TYPE	count	mean	std	min	25%	50%	75%	max
'lime mope	1.0	1.00	NaN	1.0	1.00	1.0	1.00	1.0
(ceme	1.0	1.00	NaN	1.0	1.00	1.0	1.00	1.0
.	2.0	1.50	0.707107	1.0	1.25	1.5	1.75	2.0
0	4.0	1.25	1.892969	0.0	0.00	0.5	1.75	4.0
00	1.0	6.00	NaN	6.0	6.00	6.0	6.00	6.0
...	...	...	...	...	...	...	...	...
yello	4.0	1.25	1.258306	0.0	0.75	1.0	1.50	3.0
yellow cab	1.0	1.00	NaN	1.0	1.00	1.0	1.00	1.0
yw	1.0	0.00	NaN	0.0	0.00	0.0	0.00	0.0
yy	1.0	2.00	NaN	2.0	2.00	2.0	2.00	2.0
omm	1.0	1.00	NaN	1.0	1.00	1.0	1.00	1.0

[2696 rows x 8 columns]

Tabla 2

Tabla estadística atributos cuantitativos

	UNIQUE_ID	COLLISION_ID	VEHICLE_YEAR	VEHICLE_OCCUPANTS
count	4.165923e+06	4.165923e+06	2.268633e+06	2.385526e+06
mean	1.653864e+07	3.174021e+06	2.015118e+03	8.860099e+02
std	3.346222e+06	1.497099e+06	1.482648e+02	9.097885e+05
min	1.117110e+05	2.200000e+01	1.000000e+03	0.000000e+00
25%	1.455649e+07	3.160936e+06	2.008000e+03	1.000000e+00
50%	1.754060e+07	3.682424e+06	2.014000e+03	1.000000e+00
75%	1.910157e+07	4.199720e+06	2.017000e+03	1.000000e+00
max	2.062355e+07	4.712514e+06	2.006300e+04	1.000000e+09

Tabla 3

Tabla de frecuencia por tipo de vehículo

Sedan	1040743
Station Wagon/Sport Utility Vehicle	829748
PASSENGER VEHICLE	770753
SPORT UTILITY / STATION WAGON	337927
UNKNOWN	105463
	...
12 fe	1
NTTRL	1
mac 1	1
Tlr	1
white truc	1

Tabla 4

Tabla de frecuencia del género de los conductores por tipo de vehículo

DRIVER_SEX	F	M	U
VEHICLE_TYPE			
(ceme	0	1	0
.	0	1	1
0	0	1	0
00	0	1	0
013	0	1	0
...	..	..	..
work van	0	2	0
yello	0	2	0
yellow cab	0	1	0
yy	0	1	0
omm	0	1	0

[2306 rows x 3 columns]

Tabla 5

Tabla de frecuencia de la dirección hacia donde se dirigen los tipo de vehiculos

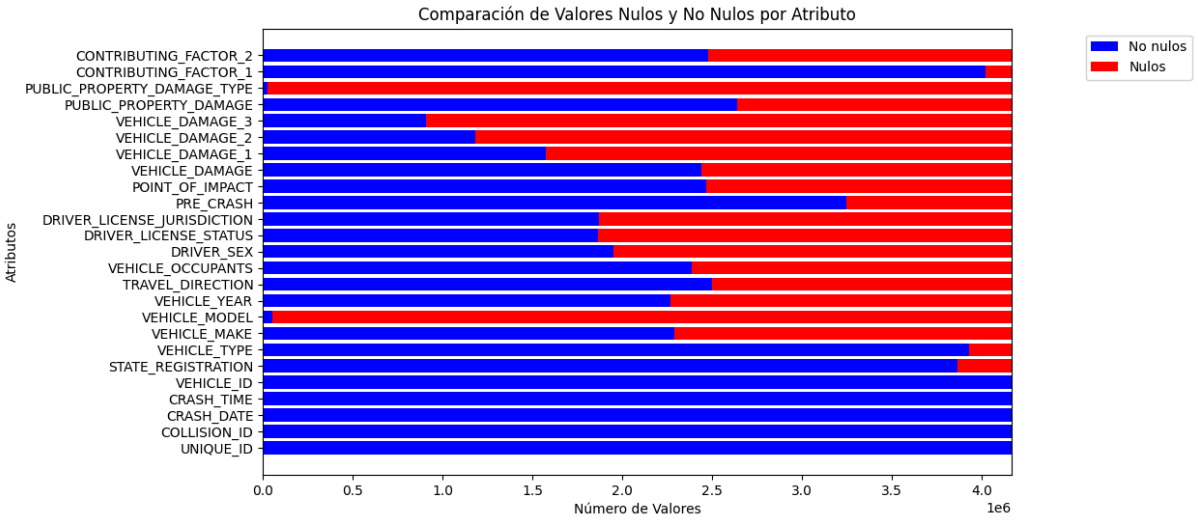
TRAVEL_DIRECTION	-	E	East	N	North	Northeast	Northwest	S	South	\
VEHICLE_TYPE										
'lime mope	0	0	0	0	1	0	0	0	0	
(ceme	0	0	0	0	0	0	0	0	1	
.	0	0	0	0	1	0	0	0	0	
0	0	0	0	0	1	0	0	0	1	
00	0	0	0	0	1	0	0	0	0	
...	..	..	...	..	...	...	...	..	...	
yello	0	0	2	0	2	0	0	0	0	
yellow cab	0	0	0	0	0	0	0	0	0	
yw	0	0	0	0	0	0	0	0	0	
yy	0	0	0	0	0	0	0	0	0	
omm	0	0	0	0	1	0	0	0	0	

TRAVEL_DIRECTION	Southeast	Southwest	U	Unknown	W	West
VEHICLE_TYPE						
'lime mope	0	0	0	0	0	0
(ceme	0	0	0	0	0	0
.	0	0	0	0	0	1
0	0	0	0	2	0	0
00	0	0	0	1	0	0
...	...	...	..	...	..	...
yello	0	0	0	0	0	0
yellow cab	0	0	0	0	0	1
yw	0	0	0	1	0	0
yy	0	1	0	0	0	0

Tabla 6

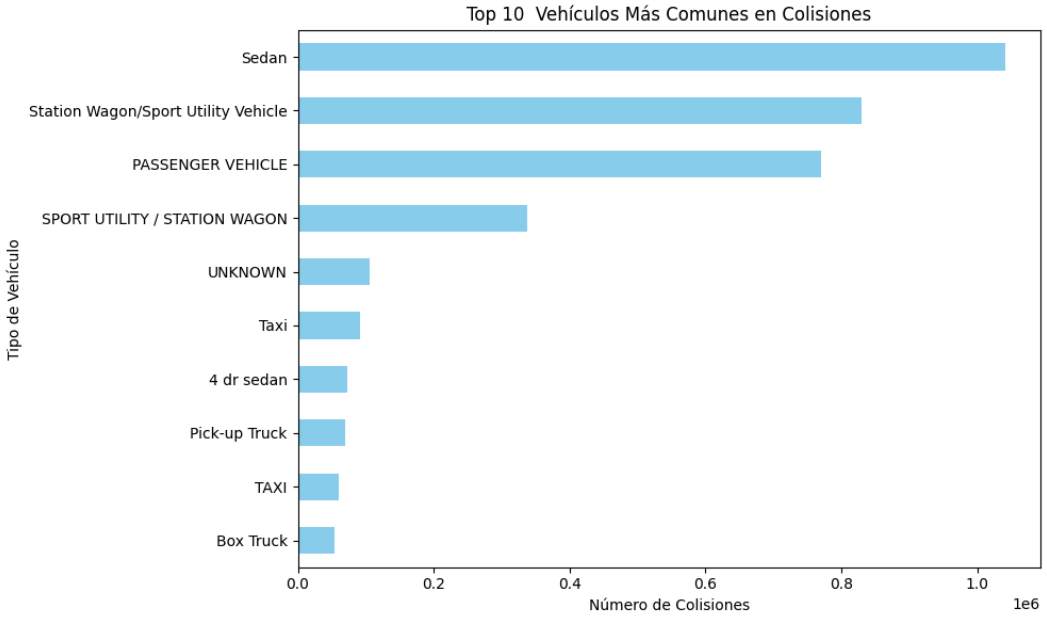
Gráficas:

Gráfico de comparación de valores nulos y no nulos en atributos



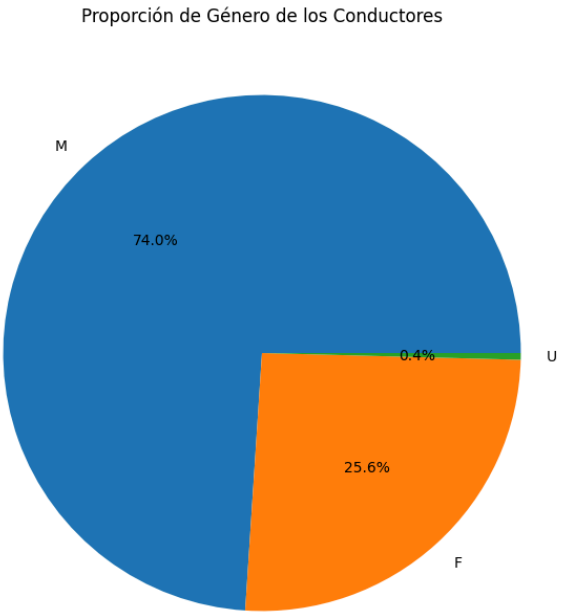
Gráfica 1

Gráfico Barras top 10 vehículos más frecuentes en las colisiones de NY



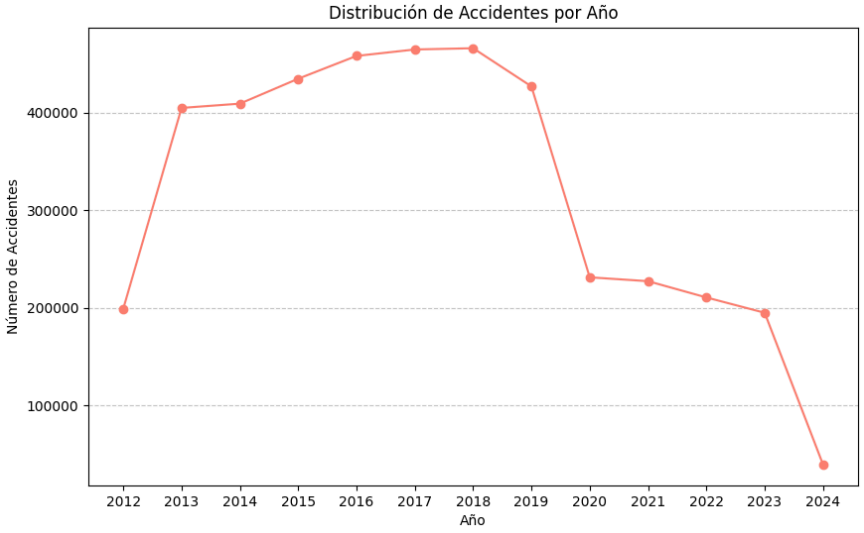
Gráfica 2

Gráfico porcentajes de géneros en los conductores de los vehículos



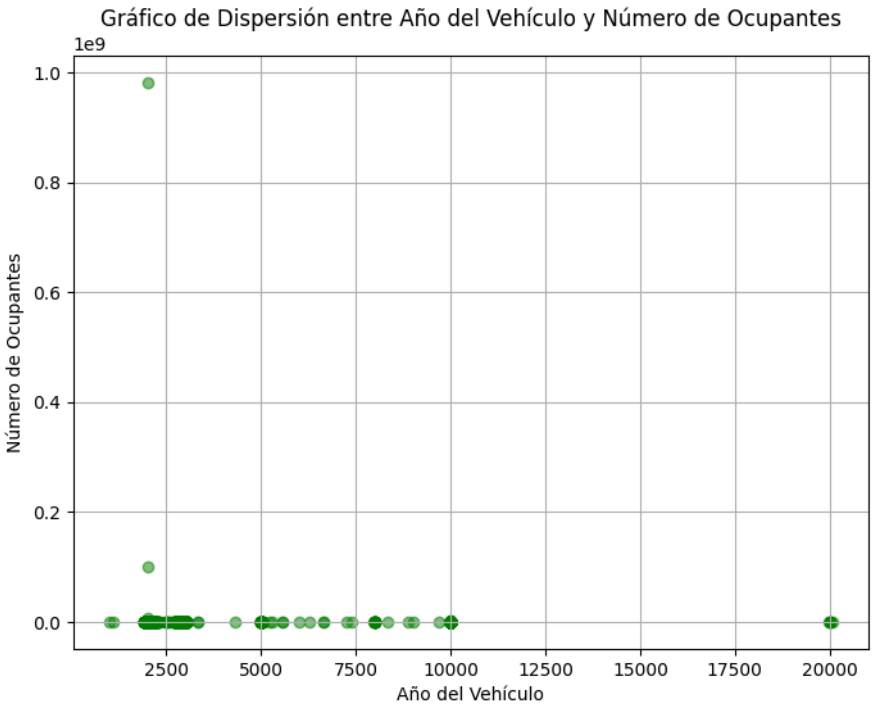
Gráfica 3

Gráfico dispersión de las colisiones desde 2012 a 2024



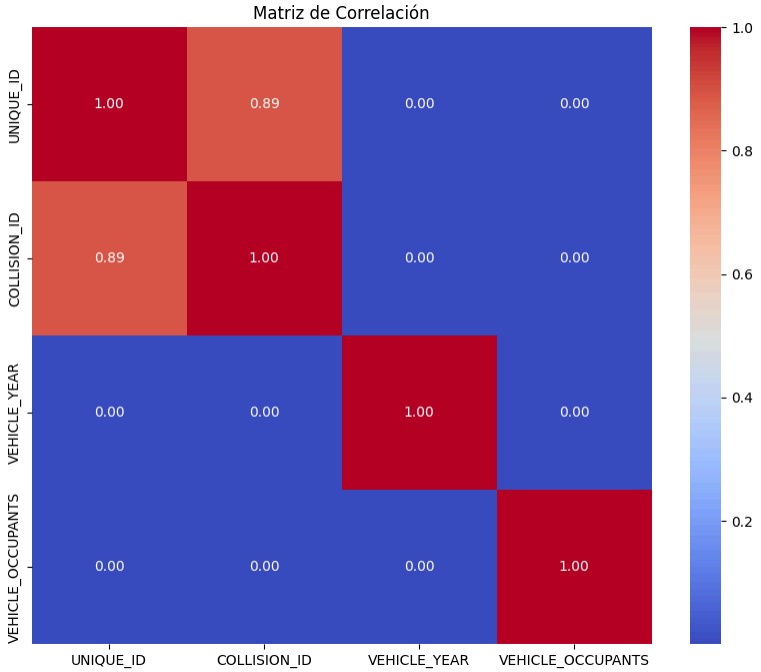
Gráfica 4

Gráfico de dispersión de número de ocupantes por año del vehículo



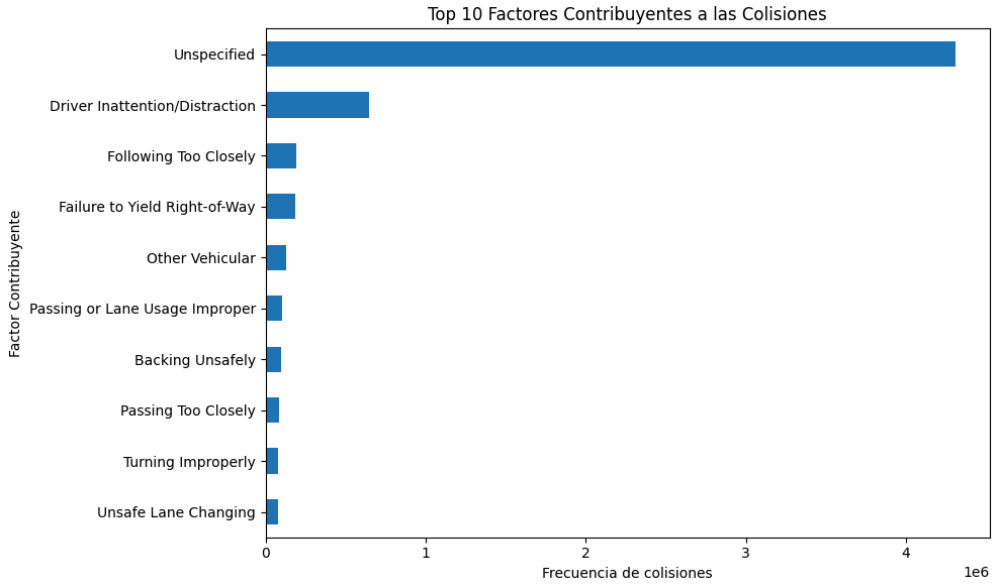
Gráfica 5

Matriz Correlación Atributos Cuantitativos



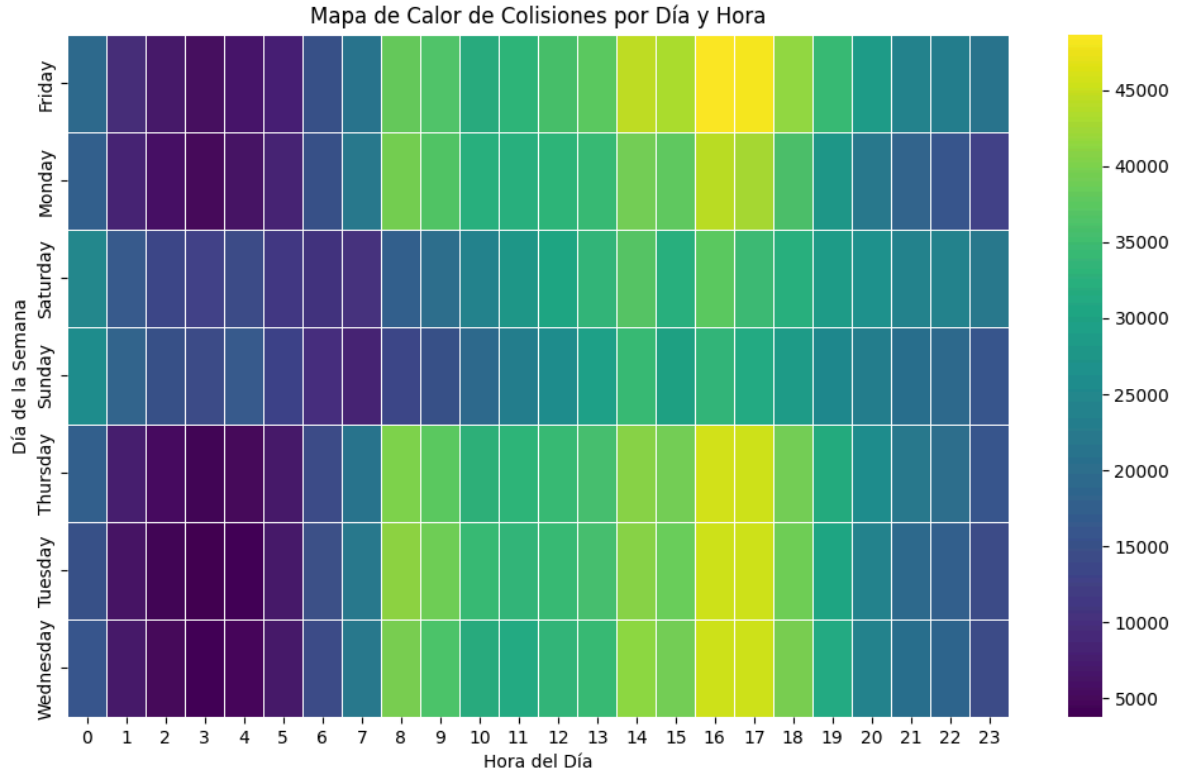
Gráfica 6

Los factores más frecuentes para que se produzca una colisión



Gráfica 7

Mapa de calor sobre frecuencia de colisiones por horas y días de la semana



Gráfica 8

Colección y Descripción de Datos:

Los datos serán cargados y procesados en el ambiente Databricks, aprovechando su capacidad para manejar y analizar grandes volúmenes de datos de manera eficiente.

Descripción de Datos:

- **Tipos de Datos:** El conjunto de datos incluye información numérica, como el año del vehículo y el número de ocupantes, así como datos de texto, como el tipo de vehículo y el estado de registro. Además, contiene datos de fecha y hora, que proporcionan información precisa sobre el momento de cada colisión.

- **Significado de Cada Atributo:**
    - UNIQUE_ID: Código único generado por el sistema para cada registro, sirve como clave primaria.
    - COLLISION_ID: Código de identificación del accidente, que se correlaciona con el ID único en la tabla de colisiones.
    - CRASH_DATE y CRASH_TIME: Fecha y hora en que ocurrió la colisión, respectivamente.
    - VEHICLE_ID: Código de identificación del vehículo asignado por el sistema.
    - STATE_REGISTRATION: Estado donde el vehículo está registrado.
    - VEHICLE_TYPE: Tipo de vehículo basado en la categoría seleccionada (por ejemplo, automóvil, bicicleta, motocicleta, etc.).
    - VEHICLE_MAKE y VEHICLE_MODEL: Marca y modelo del vehículo, respectivamente.
    - VEHICLE_YEAR: Año de fabricación del vehículo.
    - TRAVEL_DIRECTION: Dirección en la que viajaba el vehículo.
    - VEHICLE_OCCUPANTS: Número de ocupantes en el vehículo.
    - DRIVER_SEX: Género del conductor.
    - DRIVER_LICENSE_STATUS: Estado de la licencia del conductor (licencia, permiso, sin licencia).
    - DRIVER_LICENSE_JURISDICTION: Estado donde se emitió la licencia del conductor.
    - PRE_CRASH: Acción previa al choque (por ejemplo, yendo recto, girando, pasando, etc.).
    - POINT_OF_IMPACT: Ubicación en el vehículo del punto inicial de impacto.
    - VEHICLE_DAMAGE: Ubicación en el vehículo donde ocurrió la mayor parte del daño.

    - VEHICLE_DAMAGE_1, VEHICLE_DAMAGE_2, VEHICLE_DAMAGE_3: Ubicaciones adicionales de daños en el vehículo.
    - PUBLIC_PROPERTY_DAMAGE: Indica si se dañó propiedad pública (Sí o No).
    - PUBLIC_PROPERTY_DAMAGE_TYPE: Tipo de propiedad pública dañada (por ejemplo, señal, cerca, poste de luz, etc.).
    - CONTRIBUTING_FACTOR_1 y CONTRIBUTING_FACTOR_2: Factores que contribuyeron a la colisión para el vehículo designado.

- **Descripción General del Contenido:** Con más de 4 millones de registros, el conjunto de datos abarca una amplia gama de incidentes, desde colisiones menores hasta accidentes graves, proporcionando una base sólida para el análisis de la seguridad vial en la ciudad. Cada registro detalla las circunstancias específicas de un vehículo involucrado en una colisión, permitiendo un análisis profundo de los patrones y factores contribuyentes a los accidentes viales.

Reporte de Calidad de Datos

El reporte de calidad de datos es una etapa esencial en cualquier proceso de análisis, donde se evalúa la integridad y consistencia de los datos disponibles. En esta sección del documento, se presentan las técnicas propuestas para abordar valores faltantes, uno de los desafíos comunes en la gestión de datos. Estas técnicas se seleccionan cuidadosamente en función del contexto y del impacto que los datos faltantes pueden tener en los análisis posteriores.

Los resultados del conteo de datos faltantes por columna son:

UNIQUE_ID 0

COLLISION_ID 0

CRASH_DATE 0

dtype: int64

Algunas técnicas propuestas para tratar valores faltantes:

- Eliminación de filas o columnas: Si la cantidad de valores faltantes es pequeña y no afecta significativamente los análisis, podrías considerar eliminar las filas o columnas correspondientes.

- Imputación de valores: Reemplaza los valores faltantes por la media, mediana o moda de la columna según corresponda.

- Utilización de modelos de imputación: Utiliza modelos de imputación más avanzados como KNNImputer, IterativeImputer, etc., para predecir los valores faltantes basados en otros atributos.

- Codificación de valores faltantes: Si los valores faltantes tienen un significado especial, puedes codificarlos como una categoría separada.

Planteamiento de Preguntas sobre los Datos

1. ¿Hay alguna relación entre el estado de la licencia del conductor y la gravedad de las colisiones en las que están involucrados?

2. ¿Cómo ha evolucionado la cantidad de colisiones a lo largo del tiempo? ¿Hay alguna tendencia visible?

3. ¿Cómo podríamos utilizar los datos de colisiones para mejorar los algoritmos de conducción autónoma de Tesla?

4. ¿Podríamos identificar patrones estacionales en las colisiones de vehículos? Por ejemplo, ¿hay más colisiones en invierno debido a las condiciones de la carretera?

5. ¿Podríamos utilizar los datos de colisiones para informar a los conductores de Tesla sobre los factores de riesgo más comunes y proporcionar recomendaciones personalizadas para mejorar la seguridad?

6. ¿Cuál es la probabilidad de que un vehículo con daños graves en una colisión sea de cierta marca o modelo?

7. ¿Qué áreas geográficas de Nueva York tienen la mayor concentración de colisiones vehiculares, y cómo se comparan estas áreas en términos de densidad de tráfico y características urbanas?

Filtros, Limpieza y Transformación Inicial

En esta sección del documento, se aborda el enfoque para realizar las siguientes tareas, centrándose en estos tres aspectos principales: filtrado de valores atípicos, imputación de valores faltantes y eliminación de columnas con datos insuficientes. Estas acciones se llevan a cabo siguiendo criterios específicos que buscan asegurar la coherencia y la fiabilidad de los datos.

- Filtrado de valores atípicos: Eliminación de años de vehículos anómalos > 2024.

Si se observan años que están más allá de 2024 (por ejemplo, 3000 o valores negativos), podríamos considerarlos como valores atípicos o errores en los datos

- Imputación de valores faltantes: Uso de mediana y valor más frecuente.

  Si hay valores faltantes en los datos, puedes imputarlos utilizando la mediana para datos numéricos y el valor más frecuente para datos categóricos.

- Eliminación de columnas: Remoción de columnas con >50% de datos faltantes.

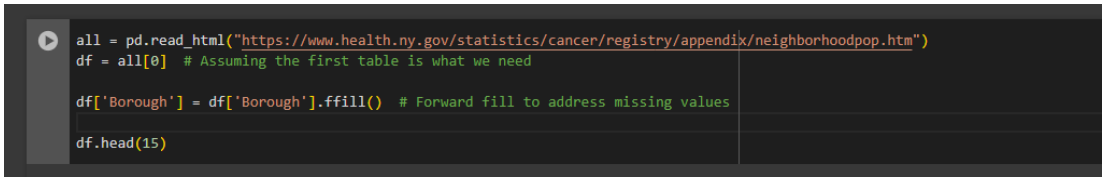
  Si las columnas en el conjunto de datos contienen más del 50% de valores faltantes, podrías considerar eliminar esas columnas, ya que pueden no proporcionar información significativa.

De esta forma y bajo estas consideraciones es que se procedió a filtrar, limpiar y transformar los datos.

BONO:

Web Scrapping

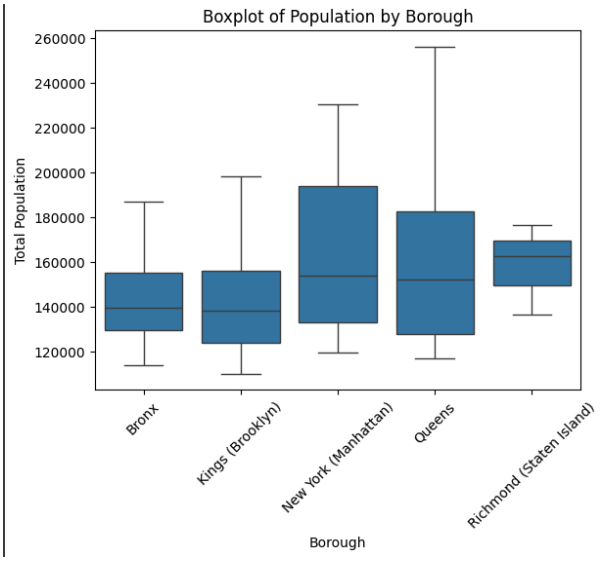
Uso la librería **pandas** para hacer el Web Scrapping.



	Borough	region	Males	Females	Total Population
0	Bronx	Riverdale, Fieldston & Kingsbridge	52133	61937	114070
1	Bronx	Wakefield, Williamsbridge & Woodlawn	65087	77848	142935
2	Bronx	Co-op City, Pelham Bay & Schuylerville	55615	65929	121544
3	Bronx	Pelham Parkway, Morris Park & Laconia	61233	67896	129130
4	Bronx	Belmont, Crotona Park East & East Tremont	75963	87740	163704
5	Bronx	Bedford Park, Fordham North & Norwood	62664	68016	130681
6	Bronx	Morris Heights, Fordham South & Mount Hope	64748	71644	136391
7	Bronx	Concourse, Highbridge & Mount Eden	67535	74968	142504
8	Bronx	Castle Hill, Clason Point & Parkchester	87605	99401	187006
9	Bronx	Hunts Point, Longwood & Melrose	80447	78645	159091
10	Kings (Brooklyn)	Greenpoint & Williamsburg	75780	76636	152416
11	Kings (Brooklyn)	Bushwick	65255	65960	131215
12	Kings (Brooklyn)	Bedford-Stuyvesant	63462	73197	136658
13	Kings (Brooklyn)	Brooklyn Heights & Fort Greene	57864	65514	123378
14	Kings (Brooklyn)	Park Slope, Carroll Gardens & Red Hook	52939	58276	111216

Gracias a la estructura del sitio no es necesario el uso de herramientas más complejas como selenium.

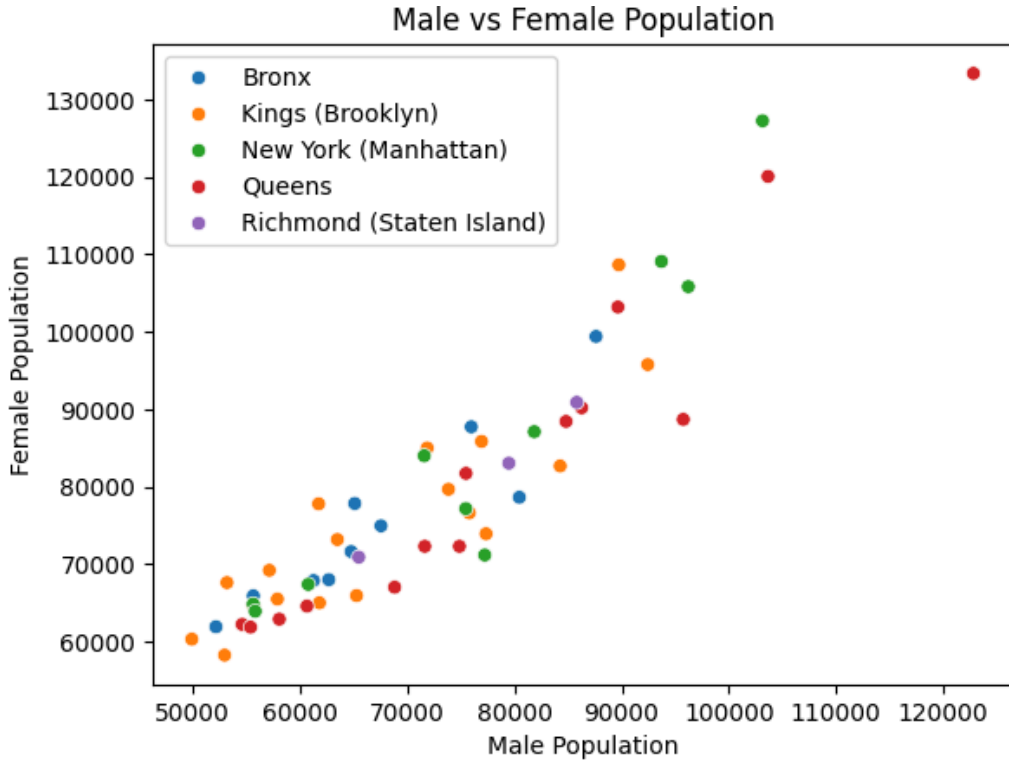
Gráficas:



Gráfica 10: Boxplot de Población por Localidad

Análisis Gráfica 10:

- **Diversidad en tamaños poblacionales:** Los boroughs de Nueva York tienen rangos de población variables, indicando diversidad en densidad habitacional.
- **Manhattan como excepción:** Manhattan tiene un rango de población más amplio y posiblemente mayor variabilidad año a año.
- **Consistencia en otros Boroughs:** Bronx, Brooklyn y Queens muestran consistencia en sus medianas poblacionales.



- **Gráfica 11: Scatter Plot Hombres vs Mujeres por Localidad**

Análisis Gráfica 11:

- **Correlación Poblacional:** Existe una correlación positiva entre la población masculina y femenina en los boroughs.
- **Equilibrio en Brooklyn:** Kings (Brooklyn) tiende a tener un equilibrio cercano entre poblaciones masculinas y femeninas.

API

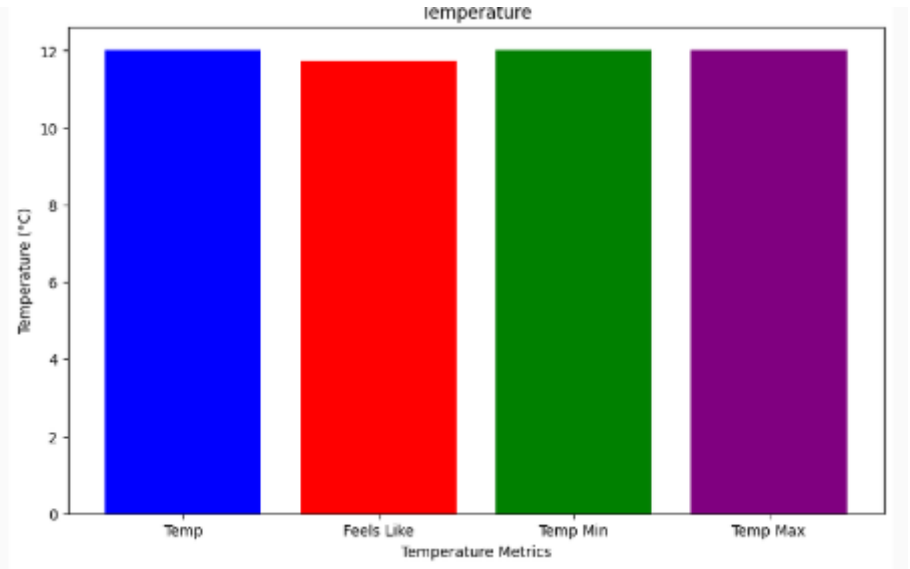
```python
import requests
import matplotlib.pyplot as plt

API_KEY = "esto NUNCA SE DEBE MOSTRAR POR SEGURIDAD, OJO CON SUS API KEYS"
LATITUDE = '4.624335'
LONGITUDE = '-74.063644'

# Ask for data in Bogota at 4/01/2024
url = f"http://api.openweathermap.org/data/2.5/weather?lat={LATITUDE}&lon={LONGITUDE}&appid={API_KEY}"
response = requests.get(url)
data = response.json()

if data.get('cod') == 200:
  print(data)
else:
  print("Error when bringing the data")
```

Fue necesario usar la versión 2.5 del API ya que la 3.0 estaba restringida.



Gráfica 12: Temperatura en el dia

La API no permitía Mostrar mucha información por llamado asi que lo más sencillo de mostrar fue la temperatura actual en Bogotá por medio de un simple bar chart.

SEGUNDA ENTREGA

Resolución de las Preguntas Problema

En este apartado del proyecto cogimos las 8 preguntas que planteamos en la primera entrega y buscamos sus soluciones, todas están interpretadas con graficos, se hace una breve explicación del gráfico y por último una solución textual a la pregunta, incluyendo los resultados del gráfico

1.¿Hay alguna relación entre el estado de la licencia del conductor y la gravedad de las colisiones en las que están involucrados?

COLLISION_SEVERITY	
Medium	2158969
Unknown	1723937
Low	276126
High	6891

Tabla. Severidad de la colisión.

En el código de esta tabla, se creó una función llamada `classify_severity` que clasifica la gravedad de las colisiones vehiculares en tres categorías: alta (`High`), media (`Medium`) y baja (`Low`), basándose en los valores de la columna `VEHICLE_DAMAGE`. La función se aplica a cada entrada en esta columna para crear una nueva columna `COLLISION_SEVERITY` en el DataFrame. Finalmente, se imprime la distribución de las diferentes categorías de gravedad para verificar la clasificación.

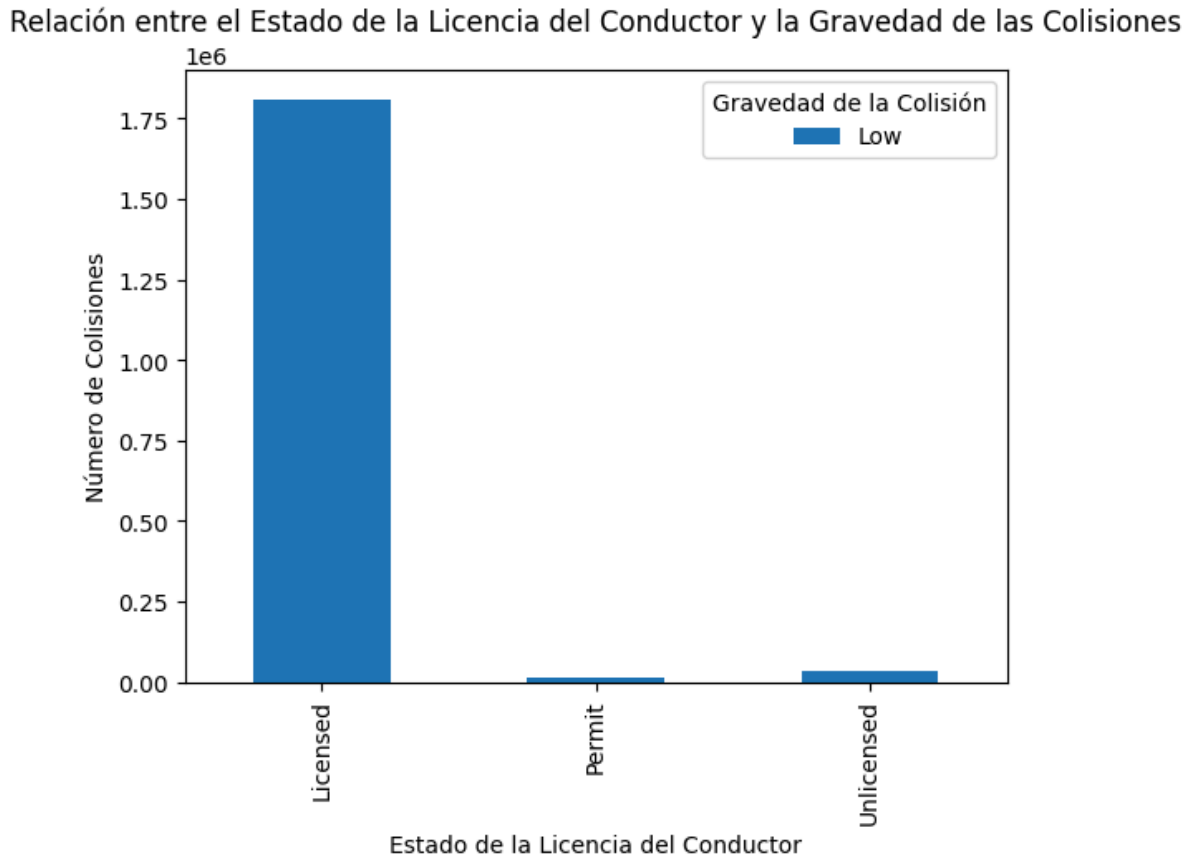


Gráfico 13. Gráfico relación estado licencia, número colisiones y gravedad

Explicación del Gráfico: El gráfico muestra la relación entre el estado de la licencia del conductor (Licensed, Permit, Unlicensed) y la gravedad de las colisiones (High, Medium, Low, Unknown). La mayoría de las colisiones involucran a conductores con licencia, con una gran parte de estas colisiones clasificadas como de gravedad media. Las colisiones con conductores con permiso o sin licencia son significativamente menores y también se clasifican principalmente como de gravedad baja.

Solución: La relación entre el estado de la licencia del conductor y la gravedad de las colisiones indica que la mayoría de las colisiones involucran a conductores con licencia completa, con una gran parte de estas colisiones siendo de gravedad media. Los conductores con permiso o sin licencia representan una fracción mínima del total de colisiones, lo que puede reflejar una menor exposición o un cumplimiento de restricciones más estrictas al conducir. Para mejorar la seguridad vial, es esencial enfocarse en estrategias de prevención de colisiones y capacitación para conductores con licencia, dado que están involucrados en la mayoría de los accidentes.

2. ¿Cómo ha evolucionado la cantidad de colisiones a lo largo del tiempo? ¿Hay alguna tendencia visible?

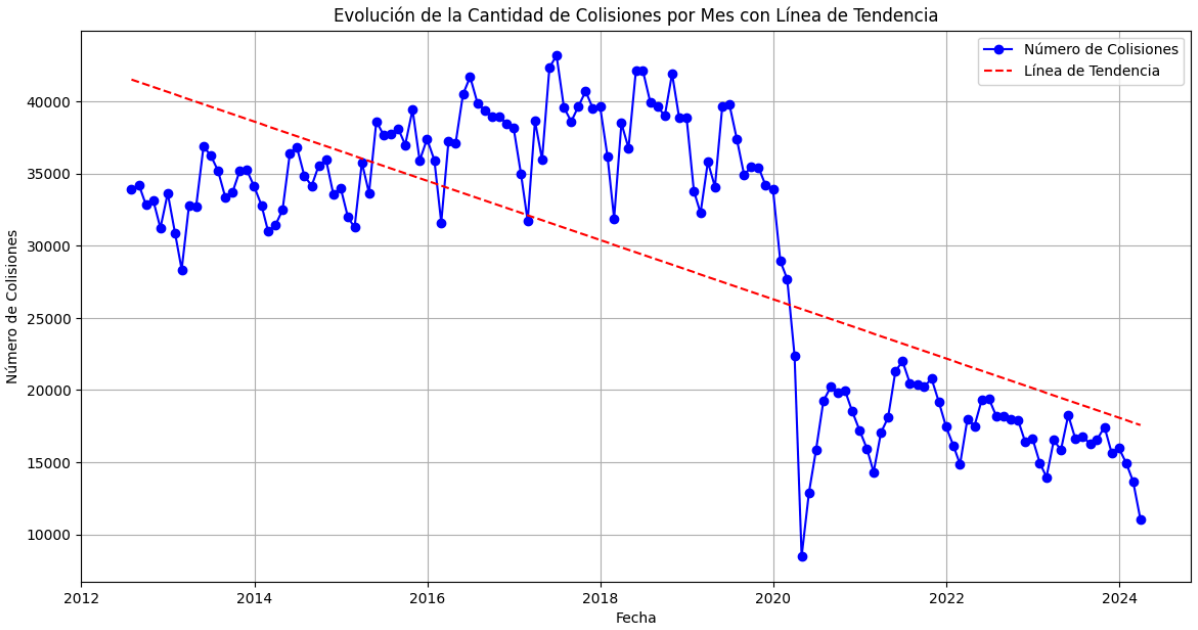


Gráfico 14. Evolución Mensual de Colisiones

Explicación del Gráfico: El gráfico muestra la evolución mensual de la cantidad de colisiones vehiculares desde 2012 hasta 2024, con una línea de tendencia superpuesta. La línea de color azul representa el número de colisiones por mes, mientras que la línea roja punteada muestra la tendencia general a lo largo del tiempo. Se observa una disminución gradual en el número de colisiones, especialmente a partir de 2020.

Solución: La cantidad de colisiones vehiculares ha disminuido a lo largo del tiempo, con una tendencia decreciente visible en el gráfico. Esta tendencia se hace más pronunciada a partir de 2020, lo que puede estar relacionado con las restricciones de movilidad debido a la pandemia de COVID-19. Para continuar con esta tendencia positiva, es crucial mantener y mejorar las medidas de seguridad vial, así como promover prácticas de conducción segura.

3. ¿Cómo podríamos utilizar los datos de colisiones para mejorar los algoritmos de conducción autónoma de Tesla?

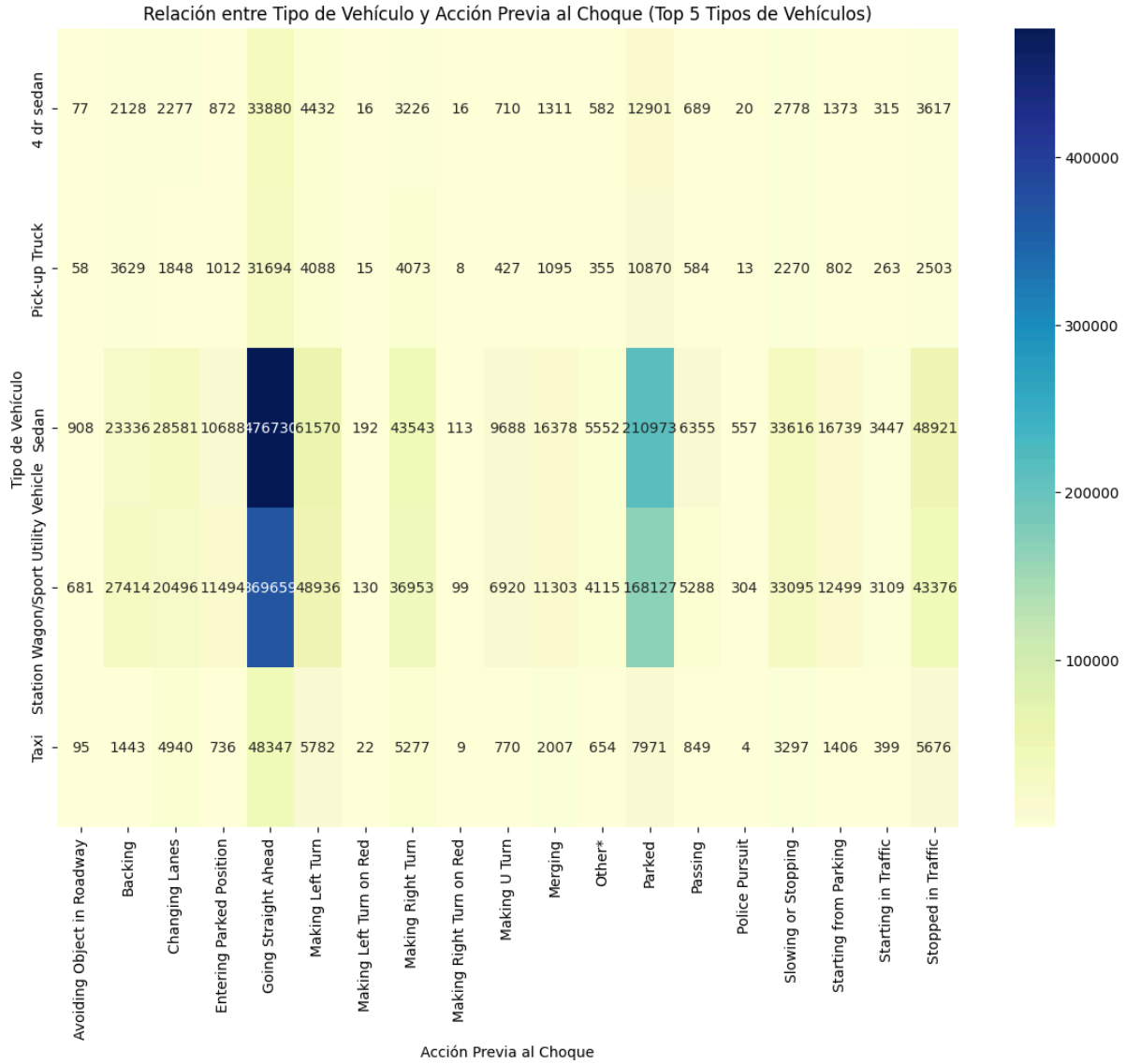


Gráfico 15. Relación Vehículo-Acción Previa

Explicación Gráfico: Este gráfico de calor muestra la relación entre los tipos de vehículos más comunes y las acciones previas al choque. Las áreas más oscuras indican una mayor frecuencia de colisiones para combinaciones específicas de tipo de vehículo y acción previa. Por ejemplo, los sedanes involucrados en colisiones mientras avanzaban en línea recta tienen una alta frecuencia.

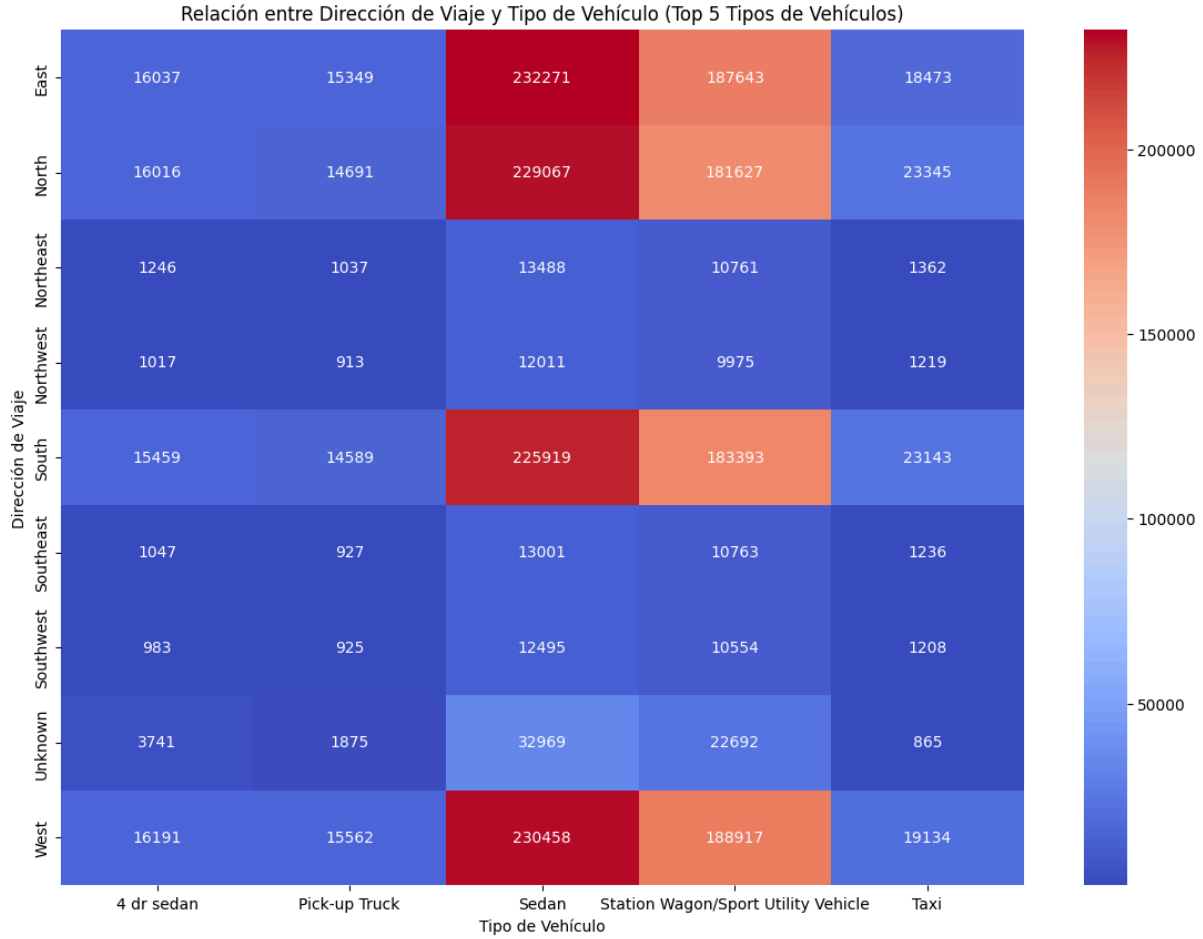


Gráfico 16. Relación Dirección-Vehículo

Explicación Gráfico: Este gráfico de calor muestra cómo se distribuyen las colisiones según la dirección de viaje y el tipo de vehículo. Las áreas en rojo indican una mayor frecuencia de colisiones. Los sedanes tienen una alta frecuencia de colisiones en varias direcciones, especialmente cuando se dirigen hacia el sur.

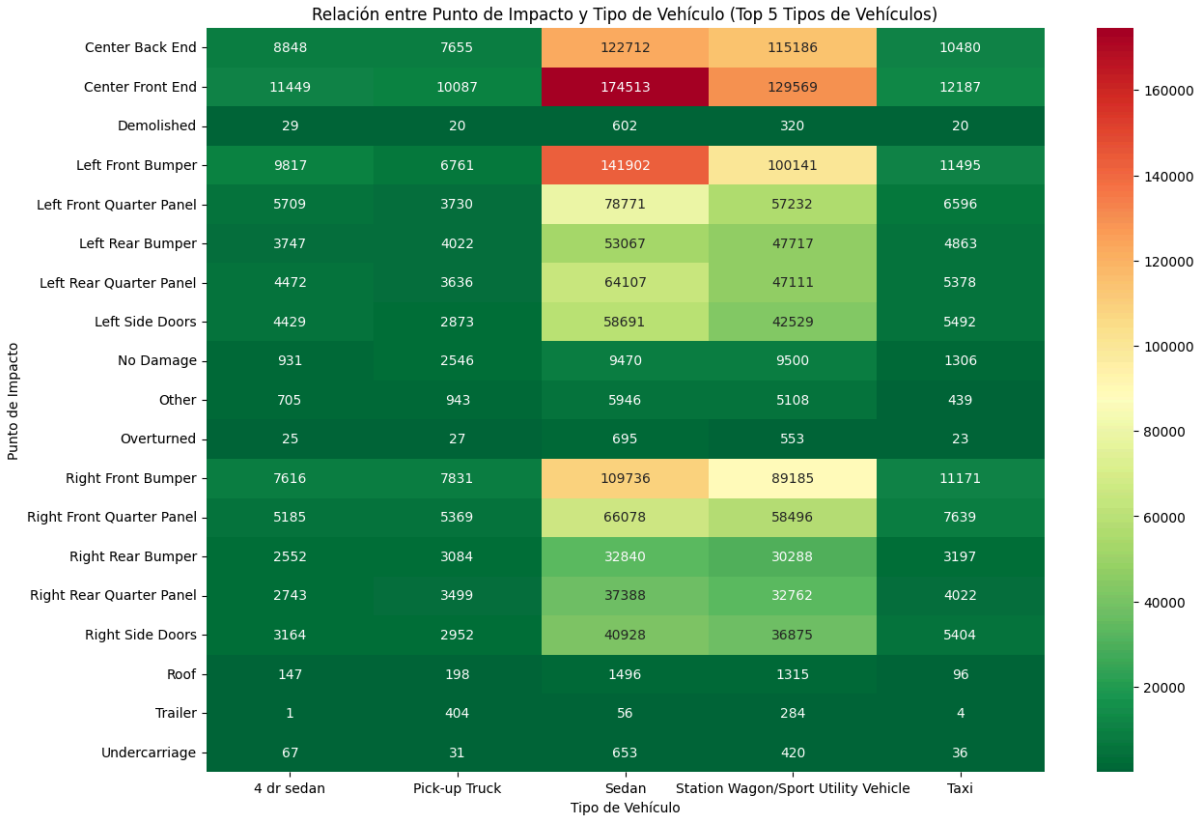


Gráfico 17. Relación Impactó-Vehículo

Explicación Gráfico: Este gráfico de calor muestra la relación entre el punto de impacto en el vehículo y el tipo de vehículo. Las áreas más oscuras indican una mayor frecuencia de colisiones en ciertos puntos de impacto. Por ejemplo, los sedanes tienen una alta frecuencia de colisiones con el impacto en el extremo frontal del vehículo.

Solución: Para mejorar los algoritmos de conducción autónoma de Tesla, podemos utilizar estos datos de colisiones para identificar patrones y situaciones de alto riesgo. Los datos permiten entrenar los algoritmos para manejar mejor las combinaciones críticas de tipo de vehículo y acciones previas, direcciones de viaje y puntos de impacto. Por ejemplo, los algoritmos pueden ajustarse para ser más cautelosos al avanzar en línea recta, particularmente en sedanes, y al abordar giros o direcciones específicas. Este enfoque proactivo en la identificación y mitigación de riesgos específicos mejorará significativamente la seguridad y eficiencia de los vehículos autónomos de Tesla.

4. ¿Podríamos identificar patrones estacionales en las colisiones de vehículos? Por ejemplo, ¿hay más colisiones en invierno debido a las condiciones de la carretera?

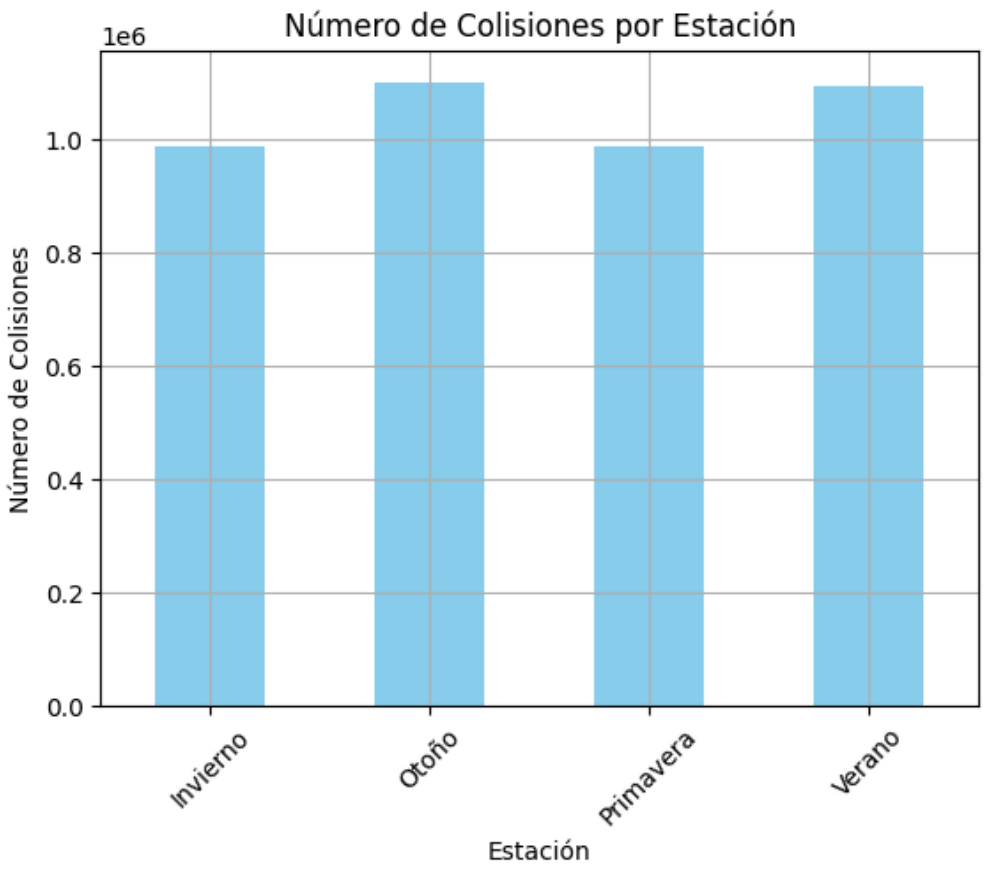


Gráfico 18. Colisiones por Estación

Explicación Gráfico: El gráfico muestra el número total de colisiones vehiculares dividido por estaciones del año: invierno, primavera, verano y otoño. Aunque hay ligeras variaciones entre las estaciones, los datos no indican un aumento significativo de colisiones en invierno en comparación con otras estaciones. El otoño parece tener un número ligeramente mayor de colisiones.

Solución: No se identifican patrones estacionales significativos en las colisiones de vehículos que indiquen un aumento claro en invierno debido a las condiciones de la carretera. La distribución de colisiones es relativamente uniforme a lo largo de las estaciones, con un ligero aumento en otoño. Para mejorar la seguridad, se podrían investigar más detalladamente otros factores estacionales y locales que puedan influir en las colisiones, como eventos específicos del clima, la luz del día y cambios en el tráfico estacional.

5. ¿Podríamos utilizar los datos de colisiones para informar a los conductores de Tesla sobre los factores de riesgo más comunes y en que horario pasan?

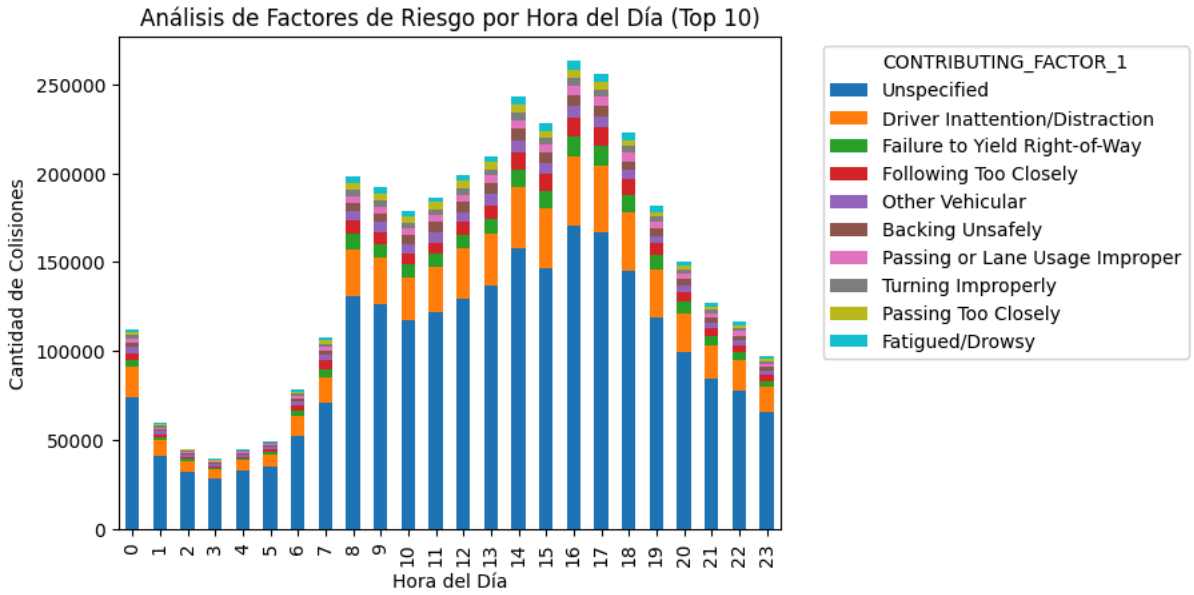


Gráfico 19. Factores de Riesgo por Hora del Día

Explicación del Gráfico: El gráfico de barras apiladas muestra la cantidad de colisiones por hora del día, destacando los factores de riesgo más comunes. Los factores de riesgo están codificados por colores, con "Unspecified" siendo el más frecuente a lo largo de todas las horas. Las colisiones tienden a aumentar en las horas pico de la mañana (7-9 am) y la tarde (4-6 pm), con factores como la distracción del conductor y la falta de ceder el paso contribuyendo significativamente.

Solución: Sí, los datos de colisiones pueden ser utilizados para informar a los conductores de Tesla sobre los factores de riesgo más comunes y los horarios en que ocurren. Al proporcionar esta información, Tesla puede alertar a los conductores sobre los momentos del día con mayor riesgo de colisión y educarlos sobre los principales factores contribuyentes, como la distracción del conductor y no ceder el paso, mejorando así la seguridad vial.

6. ¿Cuál es la probabilidad de que un vehículo con daños graves en una colisión sea de cierta marca o modelo?

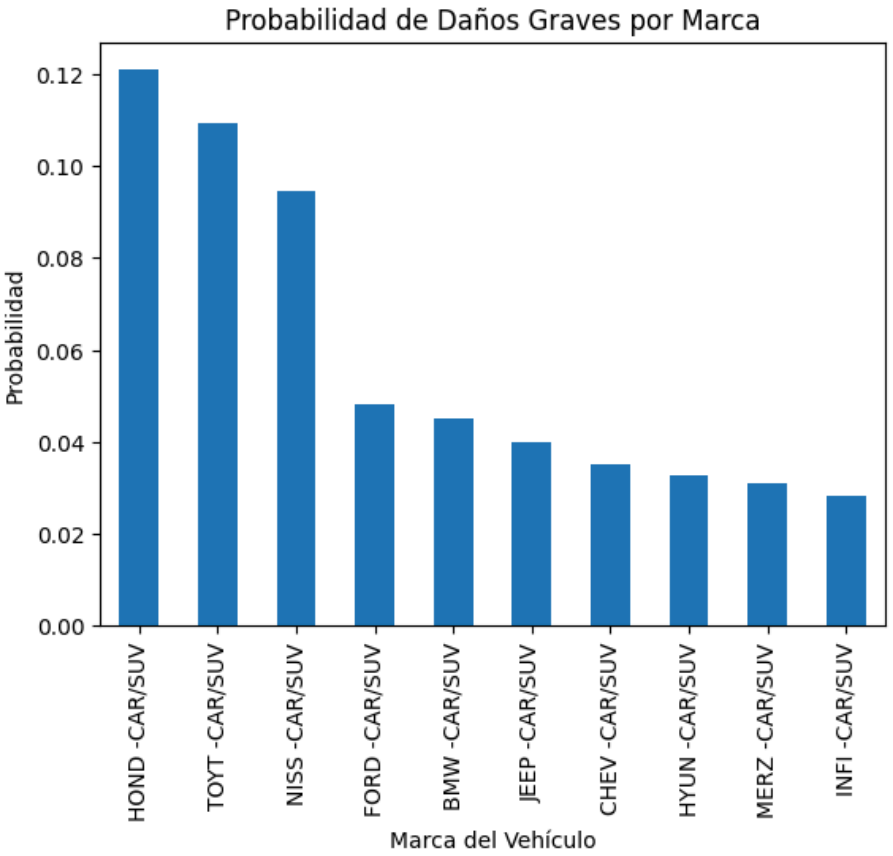


Gráfico 20. Probabilidad de Daños Graves por Marca

Explicación del Gráfico: El gráfico de barras muestra la probabilidad de que un vehículo con daños graves en una colisión pertenezca a ciertas marcas o modelos. Las barras representan la probabilidad de daños graves, con Honda (HOND) y Toyota (TOYT) mostrando las mayores probabilidades, seguidas por Nissan (NISS), Ford (FORD), y BMW. Otras marcas tienen menores probabilidades.

Solución: El análisis muestra que los vehículos de marcas como Honda y Toyota tienen una mayor probabilidad de sufrir daños graves en colisiones. Esta información puede ser utilizada para mejorar los diseños de seguridad vehicular y para alertar a los conductores de Tesla sobre los riesgos asociados con ciertas marcas y modelos, permitiendo ajustes en los algoritmos de conducción autónoma para mitigar estos riesgos.

7. ¿Cómo varía la cantidad de colisiones vehiculares según el género del conductor y la dirección de viaje del vehículo?

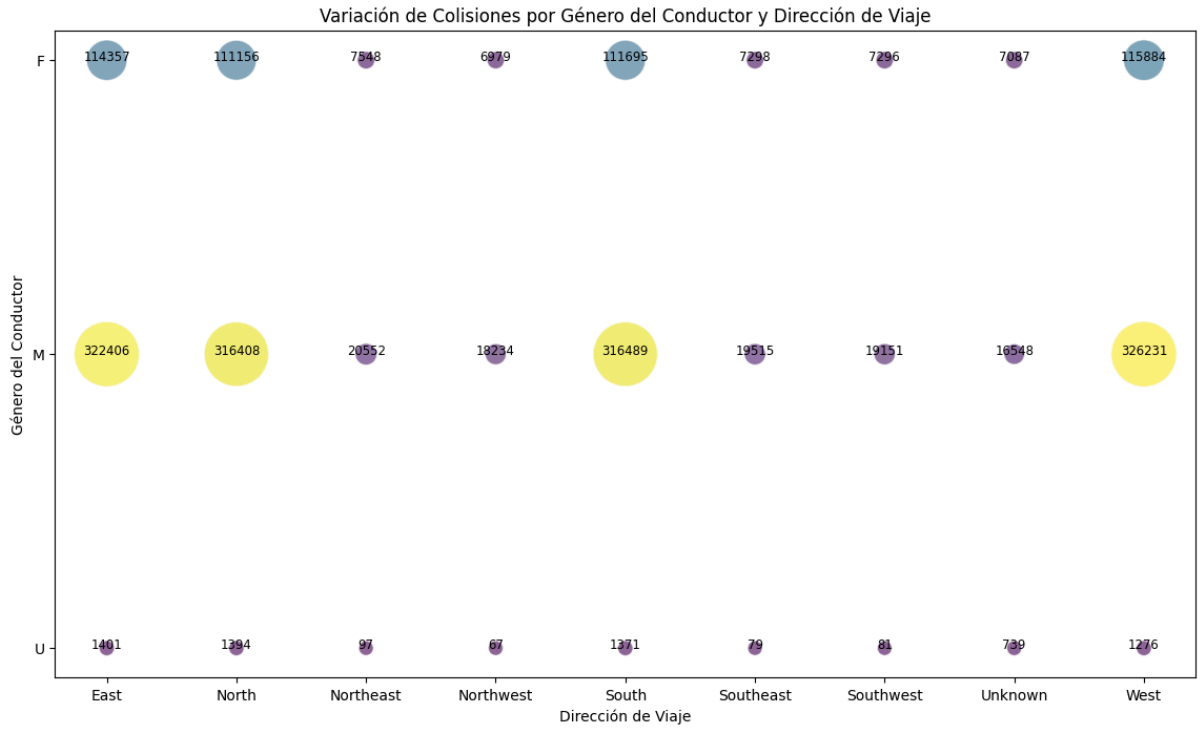


Gráfico 21. Colisiones por Género y Dirección

Explicación del Gráfico: El gráfico de burbujas muestra la cantidad de colisiones vehiculares según el género del conductor (Femenino, Masculino, Desconocido) y la dirección de viaje del vehículo (Norte, Sur, Este, Oeste, etc.). Las burbujas más grandes indican una mayor cantidad de colisiones. Se observa que los conductores masculinos tienen la mayor cantidad de colisiones en todas las direcciones, con una alta concentración en las direcciones norte y sur.

Solución: La cantidad de colisiones vehiculares varía según el género del conductor y la dirección de viaje, siendo los conductores masculinos los que presentan la mayor cantidad de colisiones en todas las direcciones. Esta información puede ser utilizada para desarrollar estrategias de seguridad vial específicas y personalizadas según el género del conductor y la dirección de viaje, mejorando así los algoritmos de conducción autónoma de Tesla para anticipar y mitigar los riesgos en función de estas variables.

8. ¿Cómo se distribuyen las colisiones vehiculares según el estado de la licencia del conductor y los factores que contribuyen a las colisiones?

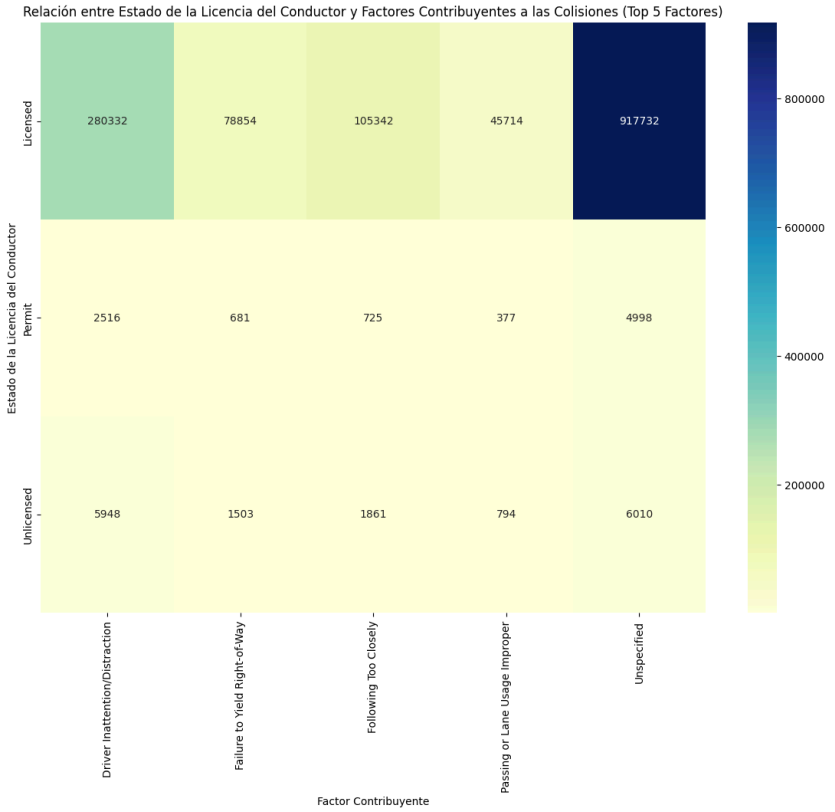


Gráfico 22. Colisiones por Estado de Licencia y Factores Contribuyentes

Explicación del Gráfico: El gráfico de calor muestra la distribución de colisiones vehiculares según el estado de la licencia del conductor (Licensed, Permit, Unlicensed) y los factores contribuyentes más comunes. Las áreas más oscuras indican una mayor frecuencia de colisiones. Los conductores con licencia completa tienen la mayor cantidad de colisiones, con la distracción del conductor y factores no especificados siendo los más comunes.

Solución: Las colisiones vehiculares están mayormente concentradas entre conductores con licencia completa, con la distracción del conductor y factores no especificados como las principales causas. Esta información es crucial para diseñar programas de concienciación y formación específicos para conductores con licencia, enfocándose en la reducción de distracciones y mejorando la atención en la carretera. También puede guiar el desarrollo de funciones avanzadas de asistencia al conductor en vehículos autónomos de Tesla.

Reporte de Calidad de Datos

El reporte de calidad de datos es una etapa esencial en cualquier proceso de análisis, donde se evalúa la integridad y consistencia de los datos disponibles. En esta sección del documento, se presentan las técnicas propuestas para abordar valores faltantes, uno de los desafíos

comunes en la gestión de datos. Estas técnicas se seleccionan cuidadosamente en función del contexto y del impacto que los datos faltantes pueden tener en los análisis posteriores.

Los resultados del conteo de datos faltantes por columna son:

Valores nulos en el DataFrame:

UNIQUE_ID	0
COLLISION_ID	0
CRASH_DATE	0
CRASH_TIME	0
VEHICLE_ID	0
STATE_REGISTRATION	4539
VEHICLE_TYPE	3460
VEHICLE_MAKE	0
VEHICLE_MODEL	1764837
VEHICLE_YEAR	10292
TRAVEL_DIRECTION	144
VEHICLE_OCCUPANTS	69
DRIVER_SEX	11
DRIVER_LICENSE_STATUS	0
DRIVER_LICENSE_JURISDICTION	17784
PRE_CRASH	4566
POINT_OF_IMPACT	76
VEHICLE_DAMAGE	113
VEHICLE_DAMAGE_1	626468
VEHICLE_DAMAGE_2	922993
VEHICLE_DAMAGE_3	1139200
PUBLIC_PROPERTY_DAMAGE	0
PUBLIC_PROPERTY_DAMAGE_TYPE	1805788
CONTRIBUTING_FACTOR_1	107

CONTRIBUTING_FACTOR_2 4151

dtype: int64

Algunas técnicas propuestas para tratar valores faltantes:

- Eliminación de filas o columnas: Si la cantidad de valores faltantes es pequeña y no afecta significativamente los análisis, podrías considerar eliminar las filas o columnas correspondientes.

- Imputación de valores: Reemplaza los valores faltantes por la media, mediana o moda de la columna según corresponda.

- Utilización de modelos de imputación: Utiliza modelos de imputación más avanzados como KNNImputer, IterativeImputer, etc., para predecir los valores faltantes basados en otros atributos.

- Codificación de valores faltantes: Si los valores faltantes tienen un significado especial, puedes codificarlos como una categoría separada.

Filtros, Limpieza y Transformación Inicial

En esta sección del documento, se aborda el enfoque para realizar las siguientes tareas, centrándose en estos tres aspectos principales: filtrado de valores atípicos, imputación de valores faltantes y eliminación de columnas con datos insuficientes. Estas acciones se llevan a cabo siguiendo criterios específicos que buscan asegurar la coherencia y la fiabilidad de los datos.

- Filtrado de valores atípicos: Eliminación de años de vehículos anómalos > 2024.

  Si se observan años que están más allá de 2024 (por ejemplo, 3000 o valores negativos), podríamos considerarlos como valores atípicos o errores en los datos

- Imputación de valores faltantes: Uso de mediana y valor más frecuente.

  Si hay valores faltantes en los datos, puedes imputarlos utilizando la mediana para datos numéricos y el valor más frecuente para datos categóricos.

- Eliminación de columnas: Remoción de columnas con >50% de datos faltantes.

  Si las columnas en el conjunto de datos contienen más del 50% de valores faltantes, podrías considerar eliminar esas columnas, ya que pueden no proporcionar información significativa.

De esta forma y bajo estas consideraciones es que se procedió a filtrar, limpiar y transformar los datos. El resultado final luego del tratamiento de datos es:

Valores nulos después de la curación de datos:

UNIQUE_ID	0
CRASH_DATE	0
CRASH_TIME	0
VEHICLE_ID	0
STATE_REGISTRATION	0
VEHICLE_TYPE	0
VEHICLE_MAKE	0
VEHICLE_MODEL	0
VEHICLE_YEAR	0
TRAVEL_DIRECTION	0
VEHICLE_OCCUPANTS	0
DRIVER_SEX	0
DRIVER_LICENSE_STATUS	0
DRIVER_LICENSE_JURISDICTION	0
PRE_CRASH	0
POINT_OF_IMPACT	0
VEHICLE_DAMAGE	0
VEHICLE_DAMAGE_1	0
VEHICLE_DAMAGE_2	0
VEHICLE_DAMAGE_3	0
PUBLIC_PROPERTY_DAMAGE	0
CONTRIBUTING_FACTOR_1	0

CONTRIBUTING_FACTOR_2 0

DAY_NIGHT 0

OCCUPANTS_CATEGORY 0

dtype: int64

Aprendizaje de Maquina:

K-Means Clustering

Justificación y Aplicación Específica:

El algoritmo K-Means se ha implementado para segmentar geográficamente los datos de colisión en la ciudad de Nueva York y clasificar tipos específicos de incidentes. Esto ha permitido identificar zonas críticas con altas tasas de accidentes, diferenciando entre áreas con colisiones leves y graves, lo cual es crucial para ajustar las medidas de seguridad y las pruebas de conducción autónoma.

Proceso Técnico Detallado:

1. Selección de Características: Se eligieron 'VEHICLE_YEAR', 'VEHICLE_OCCUPANTS', y 'CONTRIBUTING_FACTOR_1' por su potencial para revelar patrones en la gravedad y las causas de las colisiones.
2. Preprocesamiento: Las variables categóricas fueron transformadas mediante one-hot encoding para adecuarlas al modelo de K-Means, y los datos fueron estandarizados para normalizar la escala de las características numéricas, evitando así sesgos en la formación de clusters.
3. Optimización del Modelo: Utilizando el método del codo, se determinó que dos clusters son óptimos, permitiendo una clara distinción entre tipos de accidentes sin sobreajustar el modelo a los datos.

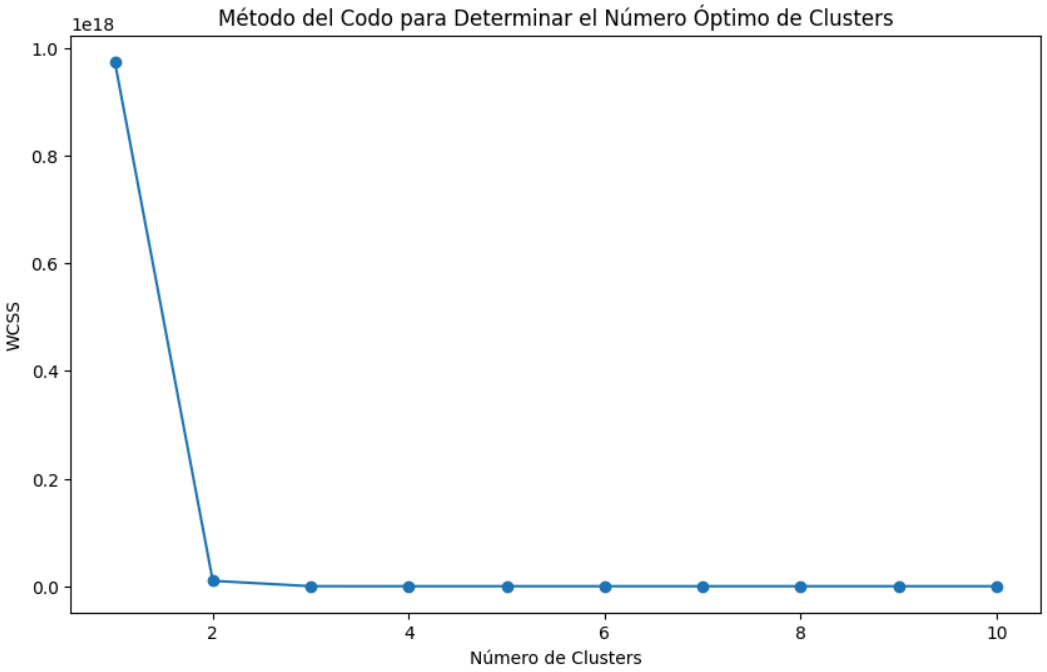


Gráfico 23. Número de Clusters Óptimo por método del codo.

Resultados y Acciones:

La agrupación reveló una concentración predominante de incidentes en un cluster, mientras que otros dos clusters fueron menos significativos, lo que sugiere una revisión en la selección y ponderación de las características para mejorar la diferenciación entre tipos de incidentes. Este análisis profundiza en la distribución y características de las colisiones, facilitando intervenciones más dirigidas y eficientes.

Árboles de Decisión

Justificación y Aplicación Específica:

Los árboles de decisión han sido fundamentales para modelar y entender cómo diferentes características de los vehículos y conductores influyen en la probabilidad de daños a la propiedad pública. Este enfoque ha proporcionado un modelo transparente y fácil de interpretar que puede ser directamente aplicado para mejorar las decisiones en tiempo real en los sistemas de asistencia al conductor.

Proceso Técnico Detallado:

1. Selección de Características: Se incluyeron datos como el tipo de vehículo, dirección de viaje, y factores contribuyentes para capturar la complejidad y diversidad de las situaciones de conducción en Nueva York.

2. División de Datos: El conjunto de datos fue dividido en un 70% para entrenamiento y un 30% para pruebas, asegurando tanto el entrenamiento robusto del modelo como una evaluación precisa de su rendimiento.
3. Entrenamiento y Limitaciones del Modelo: El modelo se entrenó con una profundidad máxima limitada para evitar el sobreajuste y garantizar que las decisiones del modelo se basen en patrones genuinos y no en anomalías de los datos.

Resumen de Desempeño del Modelo:

Categoría	Precisión	Recall	F1-Score	Soporte
No Daño (N)	0.94	1.00	0.97	192,066
Indefinido	0.82	0.00	0.00	10,636
Daño (Y)	1.00	0.00	0.00	1,284

Métricas Globales:

Métrica	Valor
Precisión Media	0.94

Recall Medio	0.33
F1-Score Medio	0.33
Exactitud	0.94

Análisis:

- Alta Precisión en 'No Daño (N)': El modelo es altamente preciso (94%) en predecir incidentes sin daño a la propiedad, con un perfecto recall de 100%, lo que significa que casi todos los verdaderos casos de 'No Daño' fueron correctamente identificados por el modelo.
- Desafíos en 'Daño' y 'Indefinido': A pesar de la alta precisión en las predicciones de daño, el recall de 0% en las categorías 'Daño (Y)' y 'Indefinido' indica que el modelo falló en identificar positivamente cualquier caso real de estas categorías.
- Necesidad de Mejoras: La baja efectividad en recall sugiere la necesidad de mejorar la capacidad del modelo para detectar y clasificar correctamente todos los casos relevantes, especialmente aquellos que involucran daño a la propiedad.

Resultados y Acciones:

La evaluación del modelo mostró un alto grado de precisión en predecir incidentes sin daño a la propiedad, pero un recall más bajo en daños a la propiedad pública, indicando áreas para mejorar la captura de datos y quizás expandir el modelo con más características o técnicas de ensamble para aumentar la sensibilidad del modelo en estos incidentes críticos.

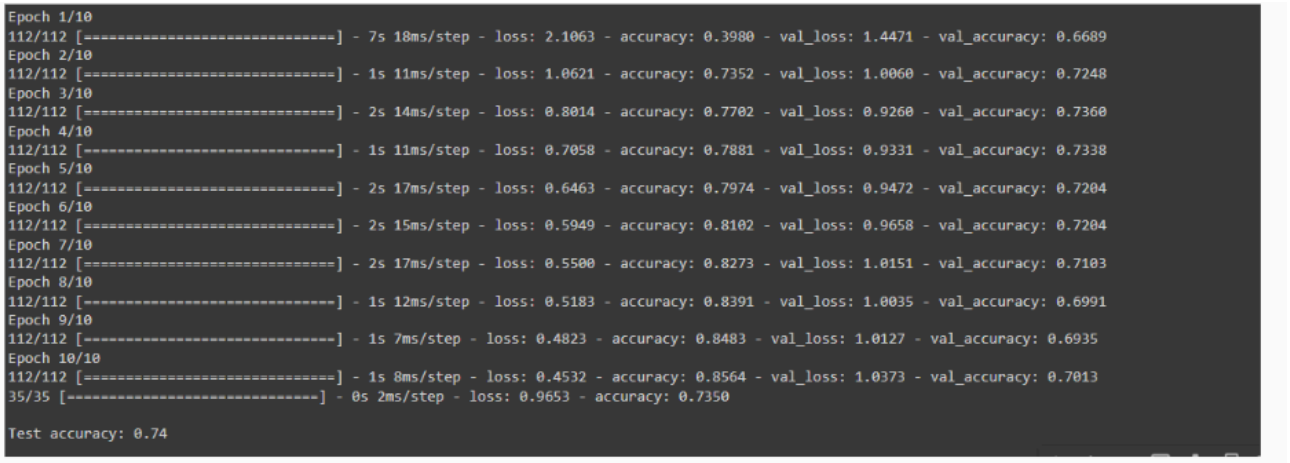
Red Neuronal con Keras

Justificación y Aplicación Específica:

Una red neuronal multicapa se utilizó para clasificar el punto de impacto en los vehículos implicados en colisiones, con el objetivo de identificar patrones específicos que puedan indicar vulnerabilidades en ciertos modelos de vehículos o situaciones de tráfico.

Proceso Técnico Detallado:

1. Selección de Características: Se centró en variables como dirección de viaje, tipo de vehículo, y punto de impacto.
2. Preprocesamiento Completo: Incluyó la codificación one-hot de variables categóricas y la estandarización de variables numéricas para preparar los datos para el entrenamiento de la red.
3. Configuración y Entrenamiento del Modelo: Se configuró la red con capas densas y funciones de activación 'relu', optimizadas con 'adam' y una función de pérdida de entropía cruzada categórica, para un clasificador multiclase.



Grafica 24: Cambio en el rendimiento del modelo en cada epoch

Resultados y Acciones:

La red alcanzó una precisión de test del 74%, un resultado prometedor para la clasificación de puntos de impacto. Los resultados sugieren áreas para futuras investigaciones en la mejora de la precisión del modelo y su aplicación en sistemas de alerta temprana para conductores y tecnologías de mitigación de colisiones en vehículos autónomos.

Conclusiones

- Los datos muestran que los impactos frontales directos son la principal causa de daños en accidentes vehiculares. Esto resalta la importancia de reforzar las medidas de seguridad en los sistemas de frenado automático de emergencia y de alerta de colisión frontal en los vehículos autónomos de Tesla.
- Los SUV y sedanes, dos de los modelos más vendidos por Tesla, aparecen desproporcionadamente involucrados en accidentes. Tesla podría aprovechar estos insights para optimizar los sistemas avanzados de asistencia al conductor específicamente para estos modelos populares.

- La identificación de las horas pico de mayor riesgo de accidentes podría permitir que Tesla ajuste dinámicamente los algoritmos de conducción autónoma para adoptar un enfoque más cauteloso y defensivo durante esos períodos críticos, minimizando así el riesgo de colisiones.

Referencias

- Data.gov Home - Data.gov. (2022). Data.gov. https://catalog.data.gov/dataset/?q=Precinct&tags=__&_tags_limit=0&_organization_limit=0&organization_type=City+Government
- Average Annual Population of NYC Neighborhoods, 2016-2020. (2016). Ny.gov. https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoodpop.htm