Trabajo Práctico Obligatorio

Ciencia de Datos

Grupo: 1

Integrantes: Papaianni Lautaro Ivan (LU: 1170805), Domingo Gabriel Ivan (LU: 1168660), Gómez Etcheber Gerónimo (LU:1125482)

Profesor: Fransisco Mariano Daniel

Tema a desarrollar: Siniestros viales en autopistas de CABA registrados por AUSA (Autopistas Urbanas Sociedad Anónima)

Consigna: Elegir un dominio a desarrollar, seleccionar la fuente de datos, adquirir un conjunto de datos desde la fuente de datos, investigar y trabajar en ese conjunto de datos y desarrollar una conclusión, análisis o predicción de los mismos.

Fuentes: https://data.buenosaires.gob.ar/dataset/seguridad-vial-autopistas-ausa , https://www.ausa.com.ar/img/mapa_de_autopistas.png

Link al Google Colab: domingo/gomez/papaianni.ipynb

Link al Looker:
https://lookerstudio.google.com/reporting/1fbe0e32-b198-4d86-b4d9-772466a82cba

Introducción:

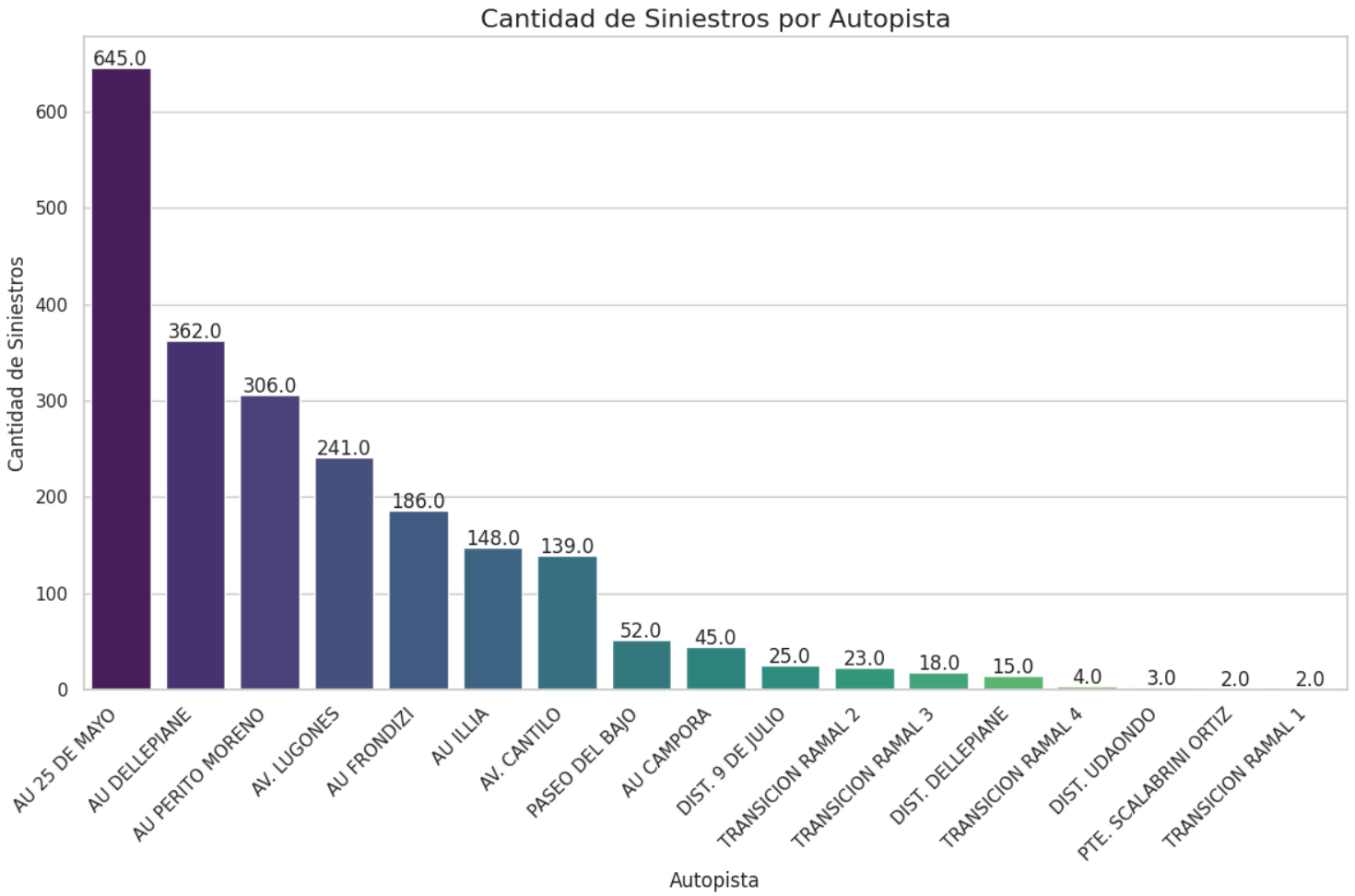
Elegimos trabajar con siniestros viales por la vasta información que puede ofrecer (debido a la frecuencialidad, lamentablemente) y por la posibilidad de poder predecir tipos de siniestros en base a la descripción, para después poder catalogarlos como leves o graves.

Análisis Exploratorio de los Datos:

De los dataset extraídos de la fuente de datos, disponemos de los siguientes tipos de datos: “AUTOPISTA;BANDA_o_RAMAL;PK;CONDICIONES_METEOROLOGICAS;SUPERFICIE_DE_LA_VIA;LESIONADOS;FALLECIDOS;TIPO_DE_SINIESTRO;MOTO;LIVIANO;BUS;CAMION;AÑO;MES;DIA;HORA;DIASEMANA;MES_STR;AÑO_STR;AÑO_MES;FECHA_COMPLETA”, de los cuales pudimos generar los siguientes gráficos relacionando los distintos tipos de datos en base a los siniestros ocurridos. Estos dataset constan de la cantidad de siniestros registrados por AUSA en los años 2022, 2023 y 2024.

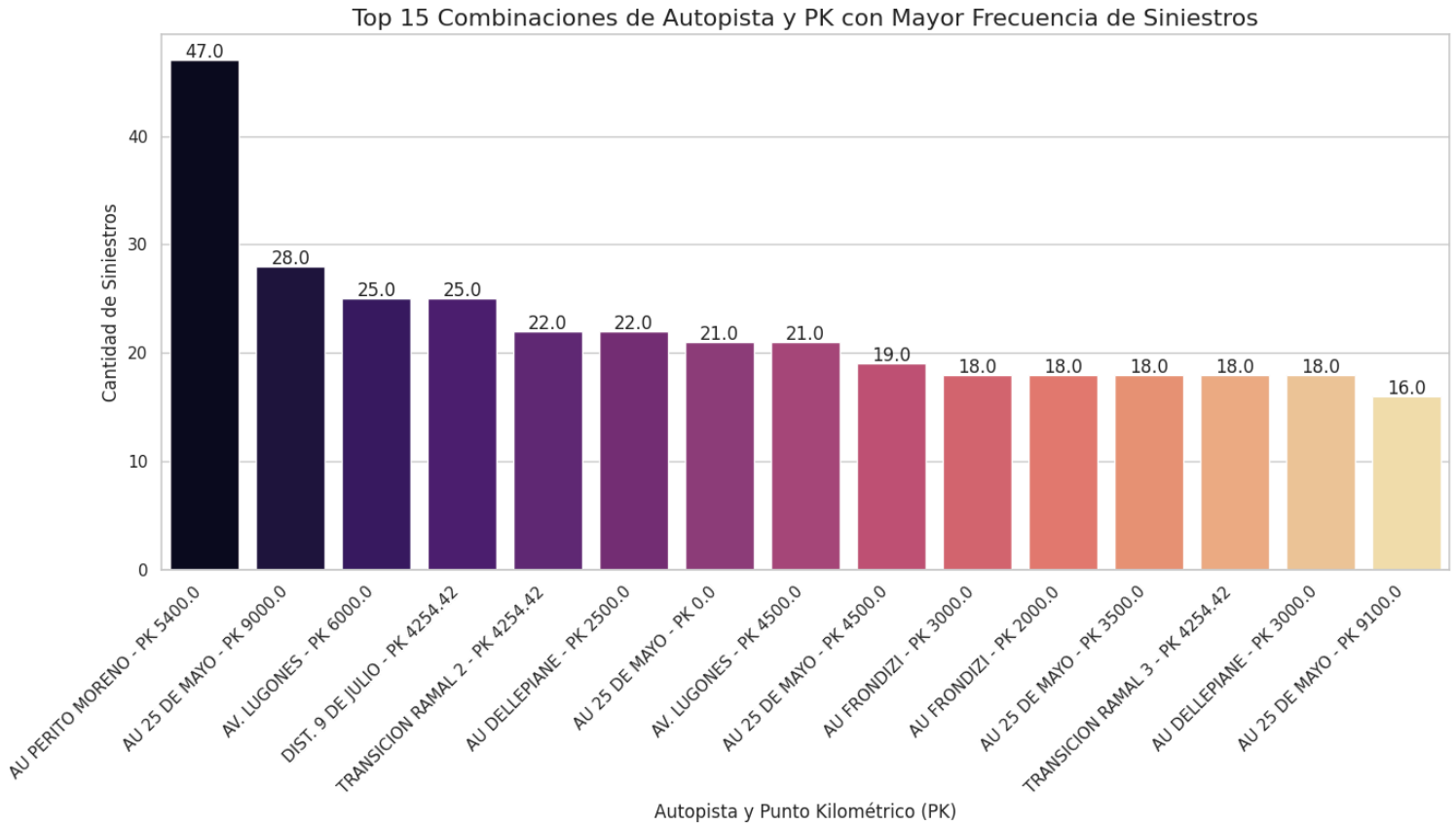
Siniestros por autopista:

Evaluamos la cantidad de siniestros ocurridos por cada autopista.



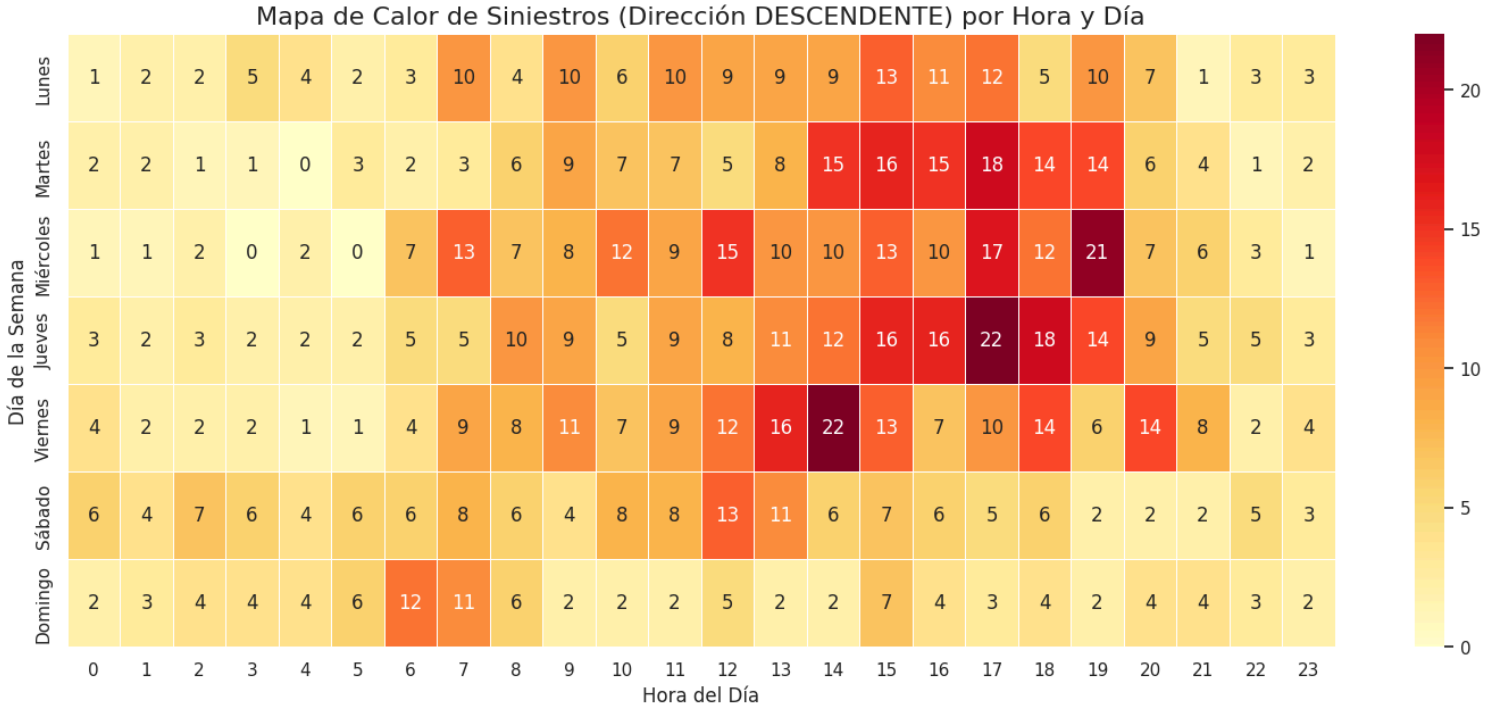
Como se observa en el histograma, la mayor cantidad de accidentes ocurrieron en la Autopista 25 de Mayo con una cantidad de 645 siniestros, seguido de la Autopista Dellepiane con 362 y Autopista Perito Moreno con 306.

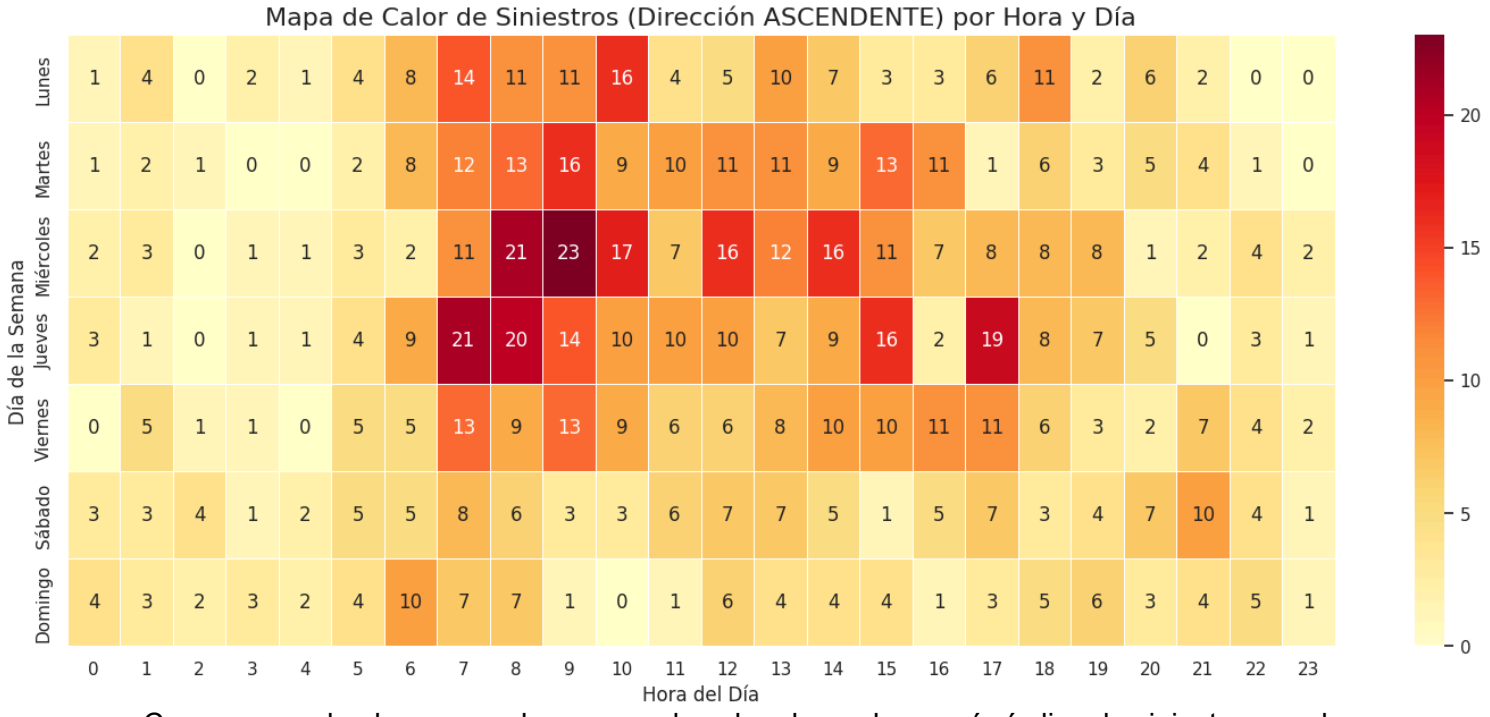
Siniestros más frecuentes por punto kilométrico y autopista:



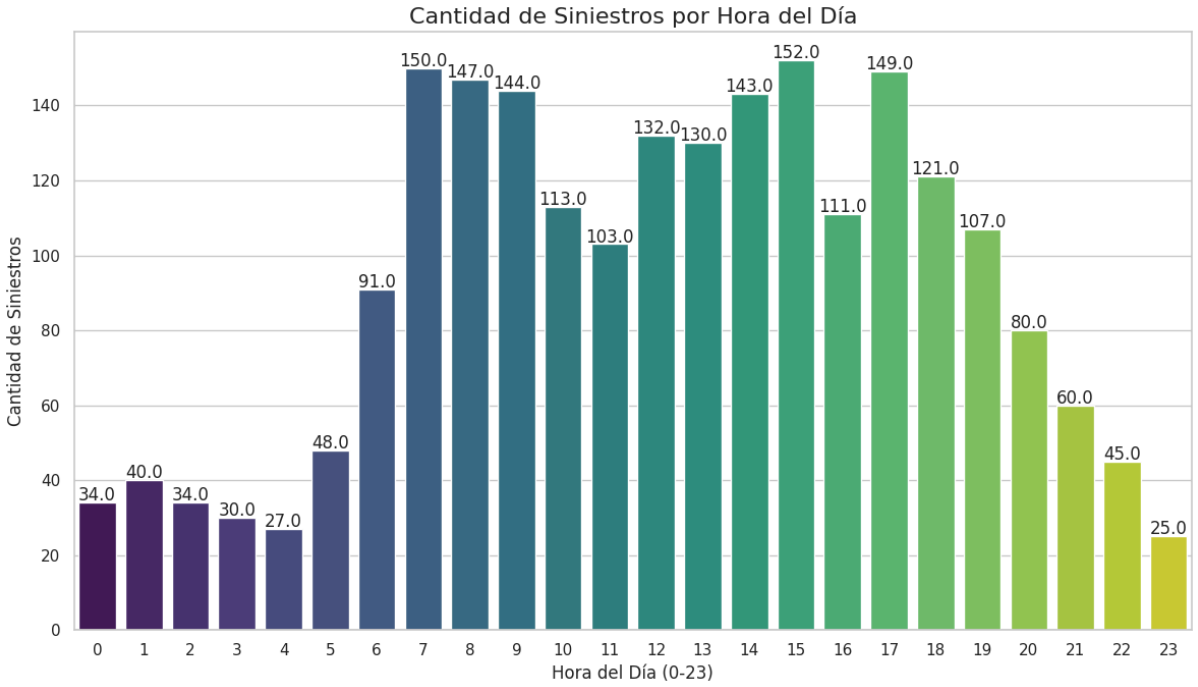
Como se observa en el histograma, la mayor cantidad de accidentes ocurrieron en la autopista Perito Moreno al kilómetro 5400 con una cantidad de 47.

Mayor cantidad de siniestros por hora dia, y ramal ascendente y descendente:



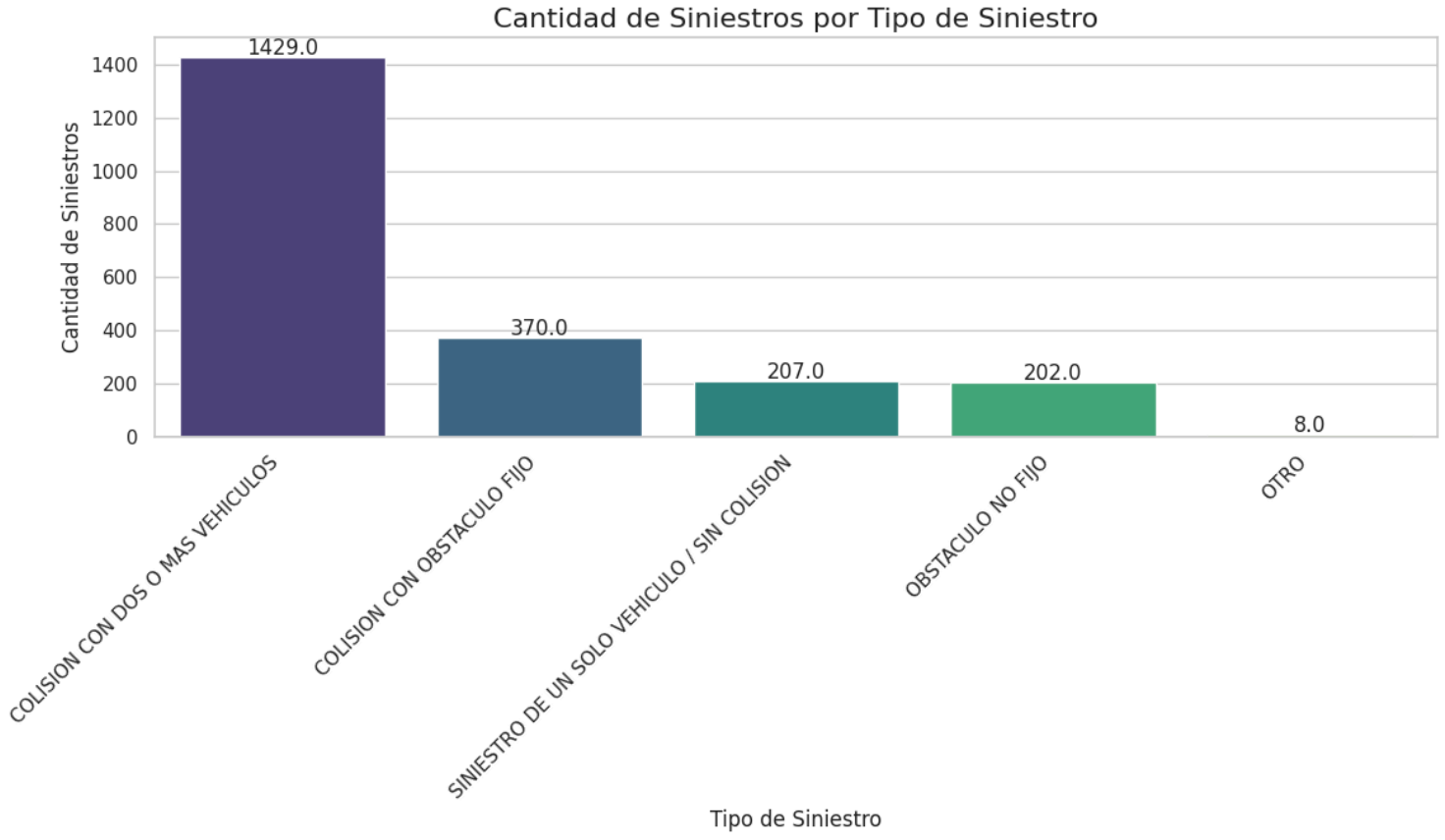


Como se puede observar en los mapas de calor el ramal con más índice de siniestros es el descendente, con la concentración entre las 12 y 19 horas entre los lunes y viernes, siendo el jueves el día con mayor cantidad de accidentes, y el jueves a la hora 17:00 y el viernes a la hora 14:00 los momentos de más accidentes. Mientras que en el ascendente, siendo el miércoles el día con más siniestros de la semana, con la concentración entre las 7:00 y 17:00 entre los lunes y viernes, con el mayor número de accidentes en el día miércoles a la hora 9:00.



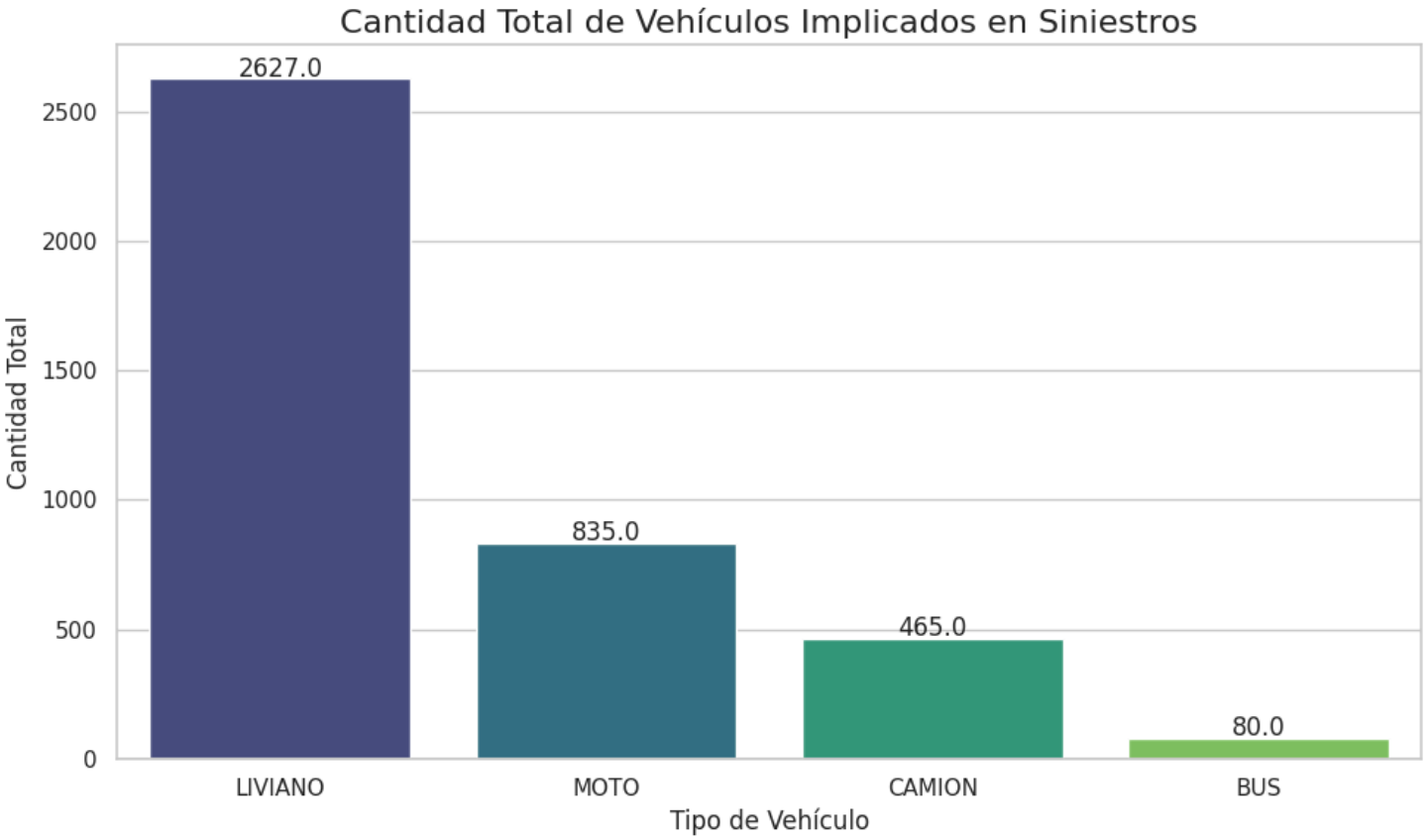
En este histograma se observa la tendencia global en la que ocurren siniestros, siendo esta entre las 7:00 y las 17:00.

Cantidad de siniestros ocurridos por tipo de siniestro:

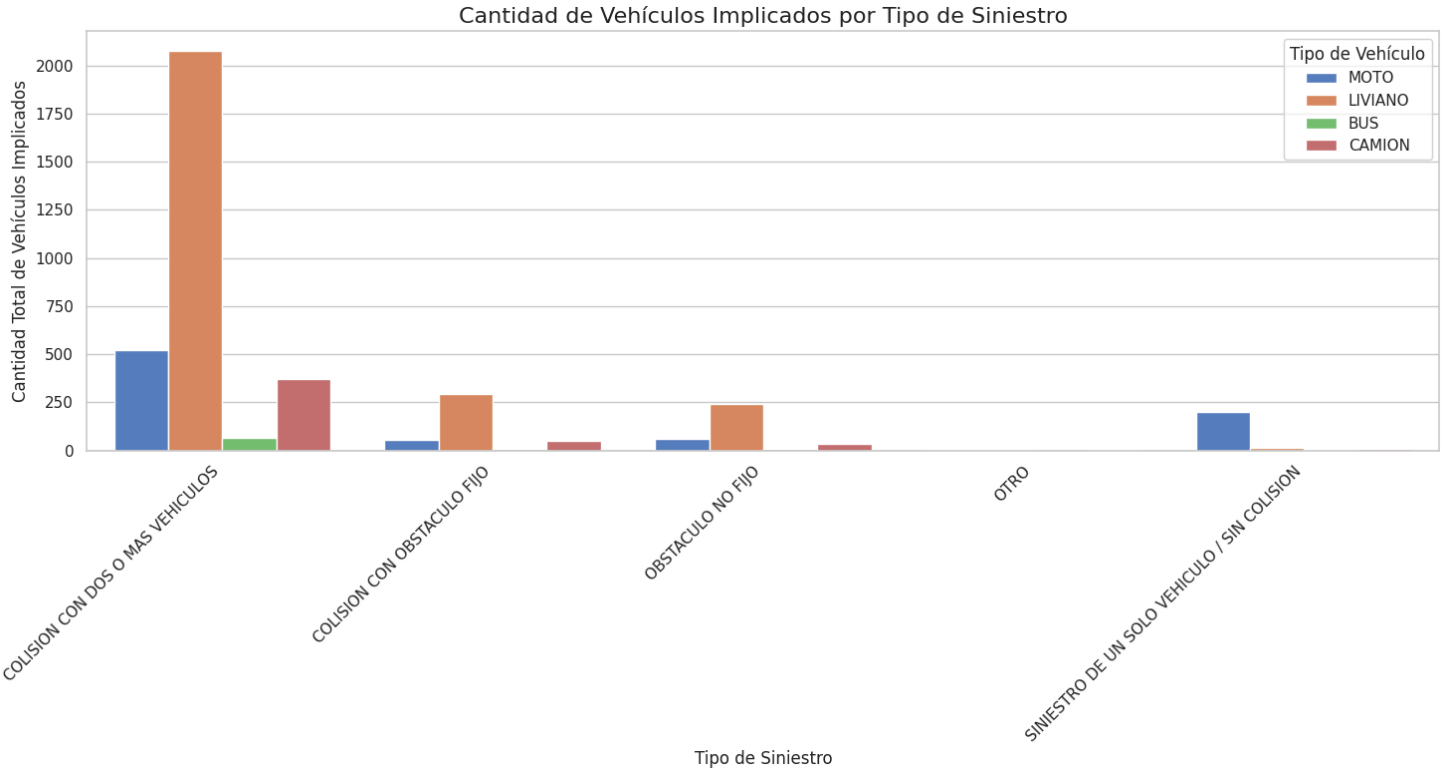


Como se puede apreciar en el histograma, la mayor cantidad de siniestros ocurridos implican colisión con dos o más vehículos, seguido de colisiones con obstáculo fijo y sin colisión o de un solo vehículo.

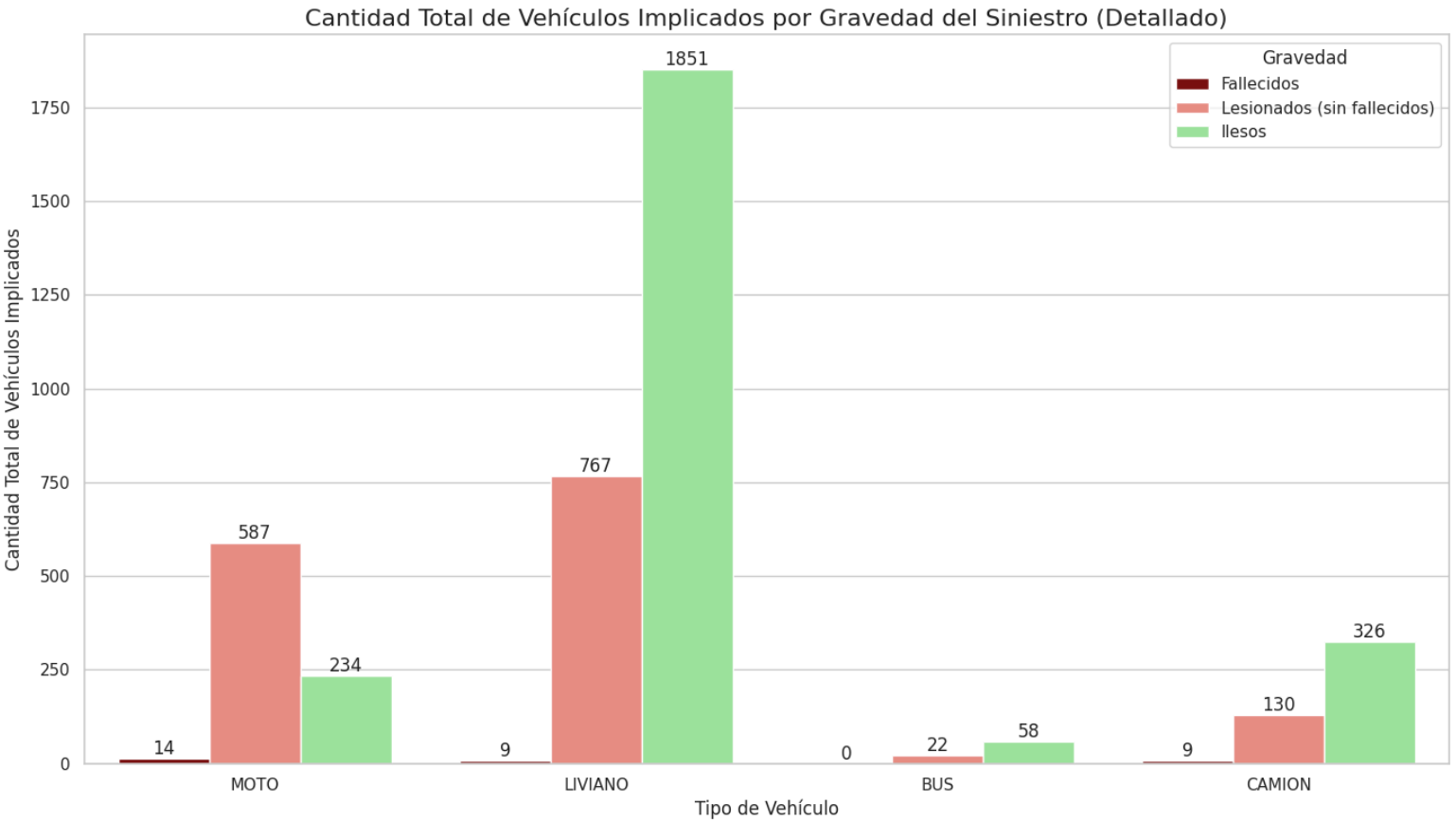
Vehículos implicados en los siniestros:



En el histograma se puede observar que en la mayoría de accidentes ocurridos, los más implicados corresponden a la categoría de los livianos.

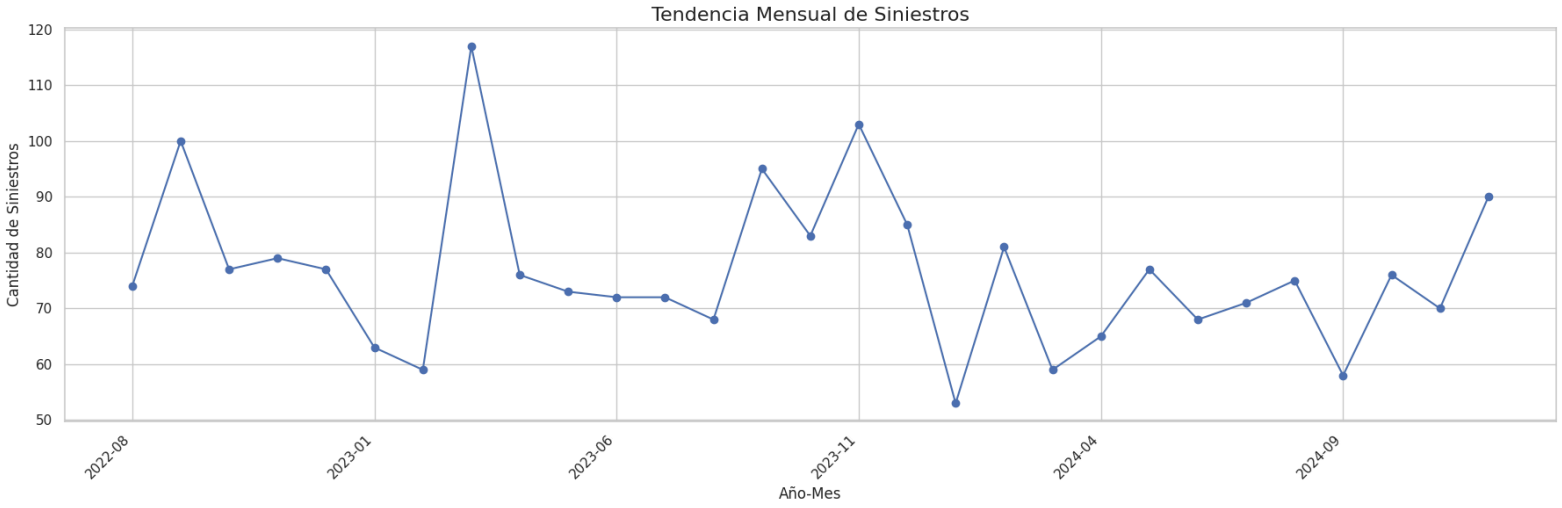


Acá se puede ver que tipo de vehículo frecuento más tipos de accidentes, siendo lo más frecuentado por todos los vehículos los accidentes de colisión entre dos o más vehículos.

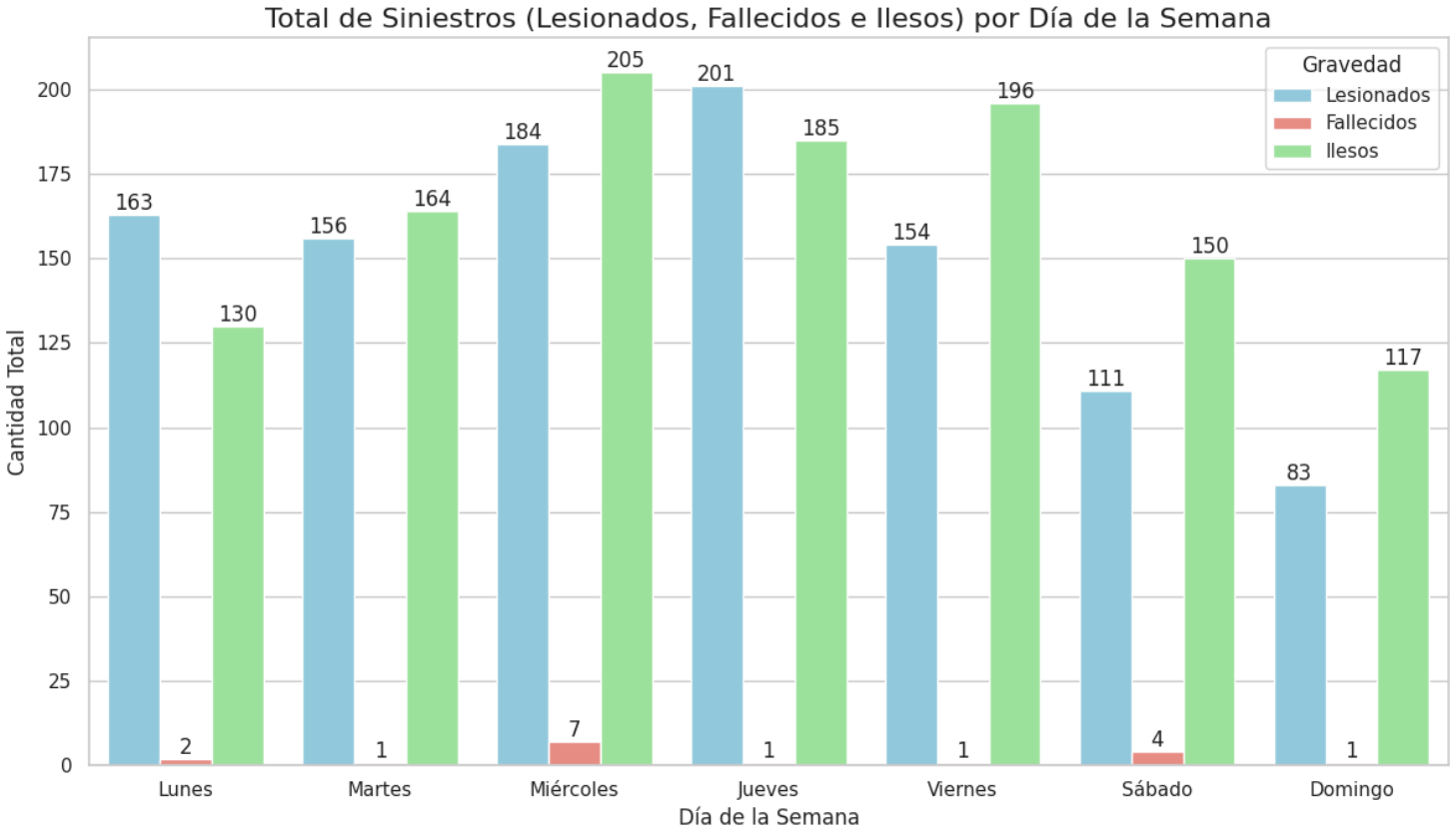


En este histograma se puede ver la cantidad de lesionados, fallecidos e ilesos por cada vehículo implicado en siniestros. Cabe aclarar que en un siniestro al estar implicados 2 tipos de vehículos distintos, los lesionados y fallecidos van a cuenta de ambos.

Siniestros por año:

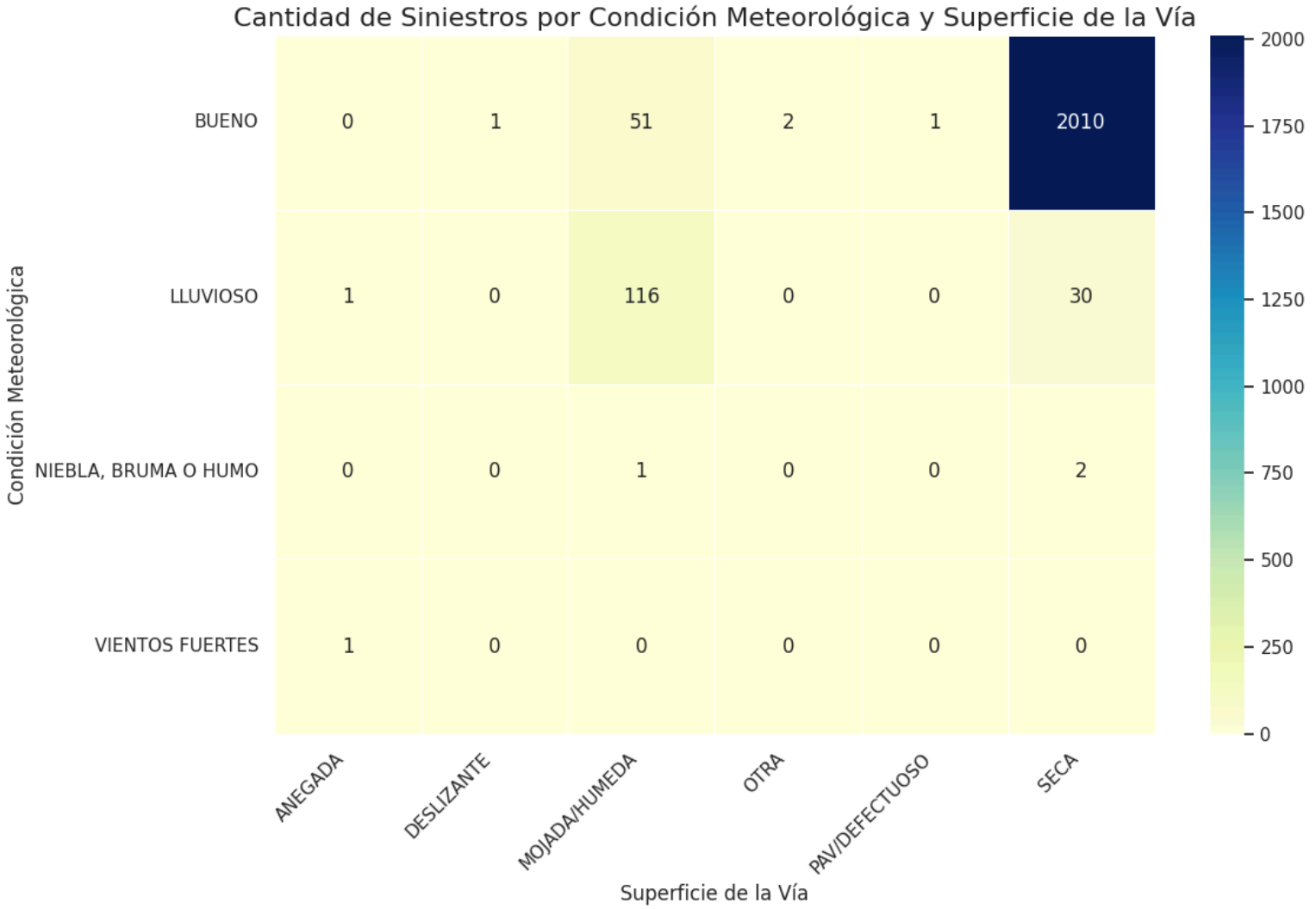


Como se ve en este gráfico de líneas, hubo una gran tendencia de siniestros por marzo del 2023, teniendo su segundo mayor pico de siniestros en noviembre del mismo año.



En este histograma se puede ver el número de lesionados, fallecidos e ilesos cada día de la semana.

Detalle de siniestros en base al estado de la vía y la condición meteorológica:



El mapa de calor muestra la cantidad de siniestros ocurridos por condición meteorológica y el estado de la superficie de la vía.

Desarrollo de algoritmo de Machine Learning:

Modelo 1: Clasificador de Gravedad de Siniestros (Grave vs Leve)

La Hipótesis y el Valor de Negocio:

- **Hipótesis:** ¿Es posible predecir, en el momento en que se reporta un siniestro, si este será "Grave" (con al menos 1 lesionado o fallecido) o "Leve" (sin víctimas), utilizando únicamente las características contextuales del evento (dónde, cuándo, tipo, vehículos)?
- **Valor de Negocio:** El objetivo principal es la **optimización de la respuesta a emergencias**. Un modelo preciso permite al centro de control de AUSA priorizar recursos, enviando potencialmente más unidades o equipos de asistencia avanzada (ej. ambulancias de alta complejidad) a incidentes que el modelo marque como "Grave", incluso antes de tener la confirmación visual.
- **Métrica Clave:** Se definió que el **Recall** para la clase "Grave" era la métrica más importante. Es preferible cometer un **Falso Positivo** (enviar recursos de más a un siniestro que resulta ser Leve) que un **Falso Negativo** (no enviar recursos suficientes a un siniestro que resulta ser Grave).

El Proceso de Construcción y Metodología:

1. **Ingeniería de Features (Target):** Creamos nuestra variable objetivo binaria (y) llamada GRAVEDAD. Se asignó 1 (Grave) si FALLECIDOS > 0 O LESIONADOS > 0, y 0 (Leve) si ambos eran 0.
2. **Análisis de Desbalance:** Constatamos un desbalance de clases moderado (63% Leves vs. 37% Graves), confirmando que la métrica "Accuracy" no sería confiable.
3. **Preprocesamiento (Pipeline):** Se construyó un Pipeline robusto de sklearn para preparar los datos (X). Este pipeline automatiza:
    - **Imputación:** Relleno de valores nulos (aunque no teníamos, es una buena práctica).
    - **Escalado:** StandardScaler para variables numéricas (como PK, HORA).
    - **Codificación:** OneHotEncoder para variables categóricas (como AUTOPISTA, DIASEMANA).
4. **Modelo Base (Random Forest):** Se entrenó un RandomForestClassifier usando class_weight='balanced' para compensar el desbalance. Este modelo arrojó un **Recall de 64%** para la clase "Grave". Si bien era un comienzo, significaba que 1 de cada 3 siniestros graves no eran detectados.
5. **Optimización Estratégica (Ajuste de Umbral):** Determinamos que, en lugar de solo re-balancear el entrenamiento, debíamos ajustar cómo el modelo *decide*. Por defecto, un modelo predice "Grave" si su confianza es $> 50\%$ (umbral 0.5).

- **Mecanismo:** Utilizamos la curva precision_recall_curve sobre los datos de entrenamiento (con validación cruzada) para encontrar un nuevo umbral.
- **Hallazgo:** Se descubrió que bajando el umbral de decisión a **0.3581**, podíamos "capturar" muchos más casos graves (aumentando el Recall) a un costo aceptable de Falsos Positivos (bajando la Precisión).

Modelo Final y Métricas Clave

- **Modelo:** RandomForestClassifier con un **umbral de decisión personalizado de 0.3581**.
- **Evaluación (sobre el Test Set):** Las métricas en el set de prueba (datos que el modelo nunca vio) validaron nuestra estrategia.

Métrica	Modelo Base (Umbral 0.5)	Modelo Final (Umbral 0.3581)	Observación
Recall (Grave)	64%	**78%**	**¡Éxito!** Detectamos 14 puntos porcentuales más de casos graves.
Falsos Negativos	~60	**36**	**¡Reducción clave!** 24 errores graves evitados.
Precisión (Grave)	71%	72%	Se mantuvo estable, un gran resultado.
Accuracy (Global)	77%	80%	La precisión general incluso mejoró.

Conclusiones e Interpretabilidad

El modelo final es robusto y cumple el objetivo de negocio. Un análisis de `feature_importance` reveló *por qué* funciona: el modelo aprendió que la presencia de una **`MOTO`** es, por lejos, el predictor más fuerte de gravedad (21% de importancia). Otros factores clave incluyen el `PK` (Punto Kilométrico), la `HORA` y el `DIA` del siniestro.

Modelo 2: Pronóstico de Siniestralidad Diaria (Serie Temporal)

La Hipótesis y el Valor de Negocio

- **Hipótesis:** ¿Es posible pronosticar el *volumen total* de siniestros (cuántos ocurrirán) para los próximos días, basándonos en los patrones históricos?
- **Valor de Negocio:** Este modelo sirve para la **planificación de recursos y gestión operativa**. Permite a AUSA anticipar días de alta demanda (ej. prever 5 siniestros en lugar del promedio de 2) y ajustar la dotación de personal de patrullas, grúas y equipos de respuesta en la traza.
- **Métrica Clave:** Al ser un modelo de regresión (predice un número), se usan el **MAE** (Error Absoluto Medio) y el **RMSE** (Raíz del Error Cuadrático Medio). El MAE nos dice, en promedio, por cuántos siniestros se equivoca el pronóstico cada día.

El Proceso de Construcción y Metodología

1. **Transformación de Datos (Agregación):** El desafío inicial fue convertir el set de datos (donde cada fila era *un* siniestro) en una serie temporal.
    - **Mecanismo:** Se utilizó `pandas.resample('D').sum().fillna(0)`. Este paso fue crucial para agrupar los siniestros por día y, fundamentalmente, **rellenar con 0** los días en que no hubo ningún siniestro, creando una serie de tiempo continua.
2. **Ingeniería de Features (Time Series):** Para que un modelo de regresión entienda el tiempo, creamos "pistas" (features) a partir del historial:
    - **Lags (Retrasos):** `lag_1` (total de siniestros de ayer), `lag_7` (total del mismo día de la semana pasada).
    - **Ventanas Móviles:** `rolling_mean_7` (promedio de siniestros de la última semana).
    - **Features de Calendario:** `dia_de_semana`, `mes` (el día de la semana resultó ser un predictor muy fuerte).
3. **División de Datos (Cronológica):** A diferencia del Modelo 1, aquí **no se pueden mezclar los datos**. Se realizó un corte cronológico: el 80% más antiguo de los datos se usó para entrenar y el 20% más reciente (el "futuro") se usó para probar.
4. **Modelo y Entrenamiento:** Se seleccionó un `XGBRegressor` (XGBoost Regressor), un modelo de *Gradient Boosting* conocido por su alto rendimiento en datos tabulares como este. Se entrenó el modelo para predecir `N_SINIESTROS` usando los lags y las features de calendario.

Modelo Final y Métricas Clave

- **Modelo:** `XGBRegressor`.
- **Evaluación (sobre el Test Set):** La métrica principal fue la visualización del pronóstico contra los datos reales en el set de prueba.
- **Resultado Visual:** El gráfico de pronóstico (línea naranja) demostró seguir muy de cerca los valores reales (línea azul).

- **Interpretación:** El modelo **logró capturar la estacionalidad semanal** de los datos, prediciendo correctamente los valles (fines de semana, con menos siniestros) y los picos (días de semana, con más siniestros).

Conclusión:

Hubo un total de 2216 siniestros de los cuales se contó 17 fallecidos y 1052 lesionados. En la mayoría de siniestros estuvieron implicados vehículos livianos, siendo estos 1706 los siniestros que involucran livianos, y en total tienden a ser colisiones entre dos o más vehículos el motivo del siniestro. Tienden a ocurrir más accidentes entre las horas 7:00 y 17:00, y también tienden a ocurrir entre los días lunes y viernes. La mayoría de siniestros ocurrieron en la autopista 25 de Mayo, pero el punto kilométrico (PK) en el que hubo más accidentes es el 5400 perteneciente a la autopista Perito Moreno y tienden a ser mayoritariamente en el ramal descendente. El año con más siniestros fue en el año 2023, siendo el mes de Marzo en el cual sucedió la mayoría.

Con toda la información recopilada del AED, decidimos realizar un análisis predictivo para poder categorizar los siniestros en base a la descripción dada. Al ocurrir un siniestro, se analiza la descripción para predecir la gravedad del mismo. Se categoriza como “Leve” el siniestro que tiene pocas posibilidades de tener lesionados o fallecidos, caso contrario se lo categoriza como “Grave”. Esto ayuda a administrar mejor los recursos a utilizar para la investigación del siniestro y para su tratamiento.