TRABAJO PRÁCTICO ESPECIAL

Accidentes viales en autopistas de Buenos Aires - 2022

Grupo 17 - Integrantes:

Octavio Collado

Micaela Mendioroz

Simón Paravich

Introducción

A continuación se realizará un informe detallado sobre el análisis de siniestros ocurridos en autopistas de Buenos Aires en el año 2022 del mes de junio al mes de diciembre, en base al registro de siniestros proporcionado por la entidad AUSA(Autopistas Urbanas S.A) y brindado por la Cátedra.
El objetivo de este informe es brindar información útil acerca de los riesgos en estas autopistas, con el fin de que se puedan tomar decisiones informadas y diseñar políticas de seguridad vial para reducirlos.

Análisis exploratorio de los datos

Descripción de las variables

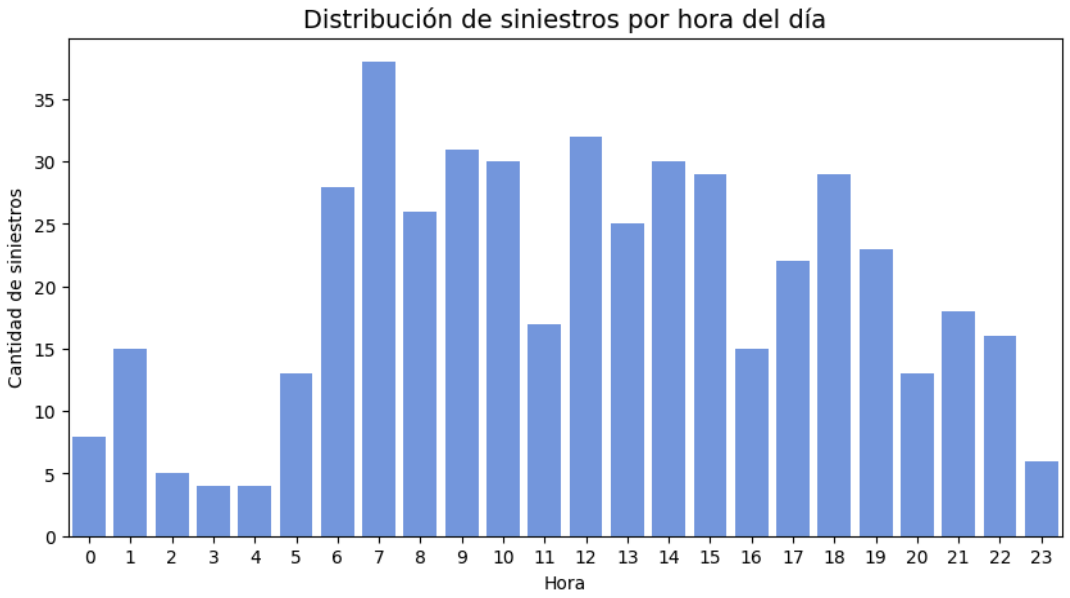
Lo primero que realizamos fue un análisis de las variables del dataset, determinando que tipo de variable es cada una y que representa.

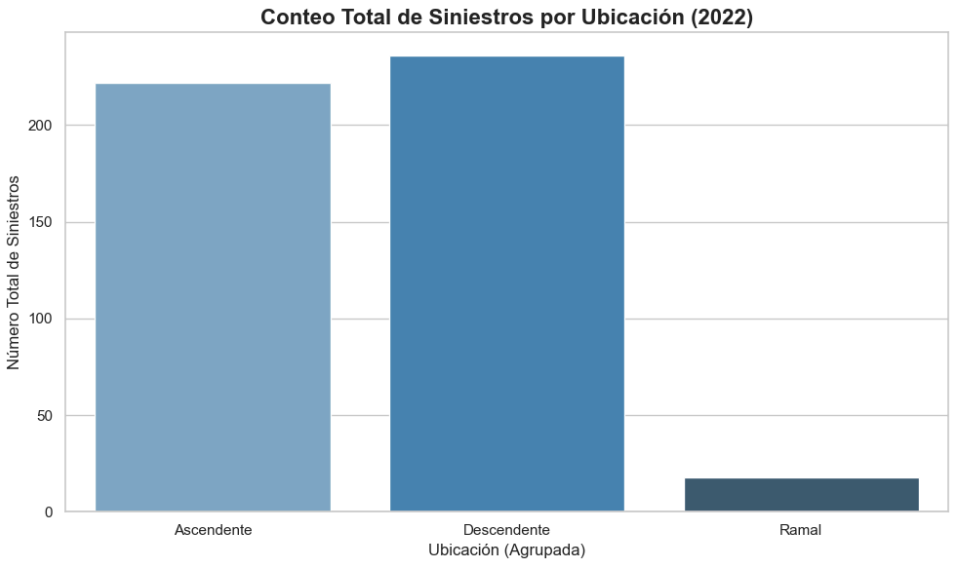
1. **Fecha**: nos da la fecha exacta(día, mes y año) en la que sucedió ese siniestro, es una variable cuantitativa discreta siendo una magnitud temporal.
2. **Hora**: describe la hora del siniestro en formato 24 horas (de 0 a 23 sin minutos ni segundos), es una cuantitativa discreta al ser una magnitud temporal.
3. **Autopista**: describe el nombre de la autopista donde ocurrió el siniestro, es cualitativa nominal.
4. **Banda y/o ramal**: indica si el siniestro ocurrió en una banda ascendente, descendente y/o ramal, el tipo de variable es cualitativa nominal.
5. **PK**: indica el punto kilométrico donde ocurrió el siniestro, es cuantitativa continua.
6. **Condiciones meteorológicas**: Describe el estado del clima al momento del siniestro(Bueno,Lluvioso,Niebla bruma o humo), es una variable cualitativa nominal.
7. **Superficie de la vía**: Describe el estado en el que estaba la superficie de la vía(Seco,mojado), es cualitativa nominal.
8. **Lesionados**: Indica el número de personas lesionadas en el siniestro, es cuantitativa discreta.
9. **Fallecidos**: Indica el número de personas fallecidas en el siniestro (si la persona falleció no se cuenta también como lesionado), es cuantitativa discreta.
10. **Tipo de siniestro**: Indica la clasificación del accidente (colisión, obstáculo no fijo,vuelco), es de tipo cualitativa nominal.
11. **Moto**: Indica la cantidad de motocicletas involucradas en el siniestro, variable cuantitativa discreta.
12. **Lviano**: Indica la cantidad de vehículos livianos involucrados, es cuantitativa discreta.
13. **Bus**: Indica la cantidad de colectivos involucrados, es cuantitativa discreta.
14. **Camión**: Indica la cantidad de camiones involucrados., es cuantitativa discreta.

Preprocesamiento de datos y primeras exploraciones

Para iniciar el análisis de siniestros viales ocurridos en las autopistas urbanas de la Ciudad de Buenos Aires reguladas por AUSA se realizó, en primer lugar, un proceso exploratorio preliminar del dataset disponible. Este paso inicial incluyó la revisión detallada de la estructura de los datos, la identificación de los tipos de variables presentes y la verificación de su coherencia interna. A partir de esta inspección se procedió a un proceso de curado y preparación de los datos, en el cual se reemplazaron valores faltantes por NAN para permitir un tratamiento uniforme, se evaluó la existencia de nombres repetidos o inconsistentes de autopistas para evitar que múltiples denominaciones que se refirieran a la misma infraestructura, y se realizaron las conversiones de tipo necesarias para facilitar el análisis estadístico y gráfico posterior. También se generaron nuevas columnas derivadas, como por ejemplo variables categóricas y variables agregadas, que aportan información útil para los estudios posteriores.

Una vez completada la preparación del dataset, se avanzó con la exploración inicial del comportamiento general de los siniestros. Entre los primeros pasos se representó la distribución de siniestros por hora del día mediante un gráfico de barras, lo cual permitió observar los momentos del día en los que se concentra la mayor actividad siniestral y detectar posibles patrones horarios. Adicionalmente, se elaboró un gráfico de conteo total de siniestros clasificados según la ubicación vial reportada en el evento (banda ascendente, banda descendente o ramal), con el objetivo de identificar si alguna de estas secciones presentaba una frecuencia marcadamente mayor o algún indicio preliminar de peligrosidad diferencial. A partir de estas observaciones descriptivas surgieron las primeras ideas sobre posibles patrones subyacentes, lo que permitió plantear hipótesis concretas para ser evaluadas posteriormente mediante métodos estadísticos formales.





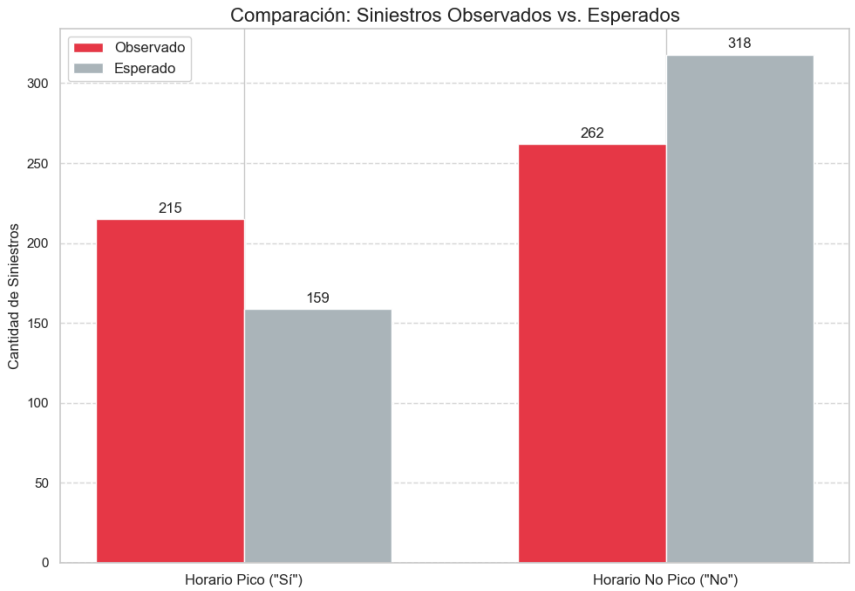
Desarrollo de las hipótesis

hipótesis 1: "Existe una mayor concentración de siniestros durante las horas pico en comparación con las horas no pico."

Para abordar la hipótesis “Existe una mayor concentración de siniestros durante las horas pico en comparación con las horas no pico”, en primer lugar se estableció una definición operativa de horario pico. Según la información provista por AUSA (https://www.ausa.com.ar/sections/tarifas.html), el horario pico corresponde, de lunes a viernes y en ambos sentidos, a las franjas comprendidas entre 07:00–11:00 h y 16:00–20:00 h. En el caso de los fines de semana (sábados, domingos y feriados), se consideran horarios pico las franjas 11:00–15:00 h en sentido provincia y 17:00–21:00 h en sentido Centro.

A partir de esta definición y utilizando las variables DIA_SEMANA (variable derivada que clasifica los días como hábiles o fines de semana) y HORA, se construyó una nueva variable categórica denominada HORA_PICO, con dos categorías: “Sí” (correspondiente al horario pico) y “No” (fuera del horario pico).

Desde el punto de vista metodológico, se determinó que una semana completa comprende 168 horas, de las cuales 56 horas (33,33%) corresponden al horario pico y 112 horas (66,67%) al horario no pico. Estas proporciones fueron utilizadas como frecuencias esperadas bajo el supuesto de que los siniestros se distribuyen de manera aleatoria y uniforme a lo largo del tiempo. Con esta información, se procedió a comparar las frecuencias observadas y esperadas mediante un gráfico de barras agrupadas, a fin de obtener una primera aproximación descriptiva.



Dado que el análisis exploratorio inicial sugirió una mayor concentración de siniestros en horario pico, se aplicó posteriormente un test de bondad de ajuste Chi-cuadrado con el objetivo de evaluar si la distribución observada difiere significativamente de la distribución esperada bajo el supuesto de aleatoriedad temporal.

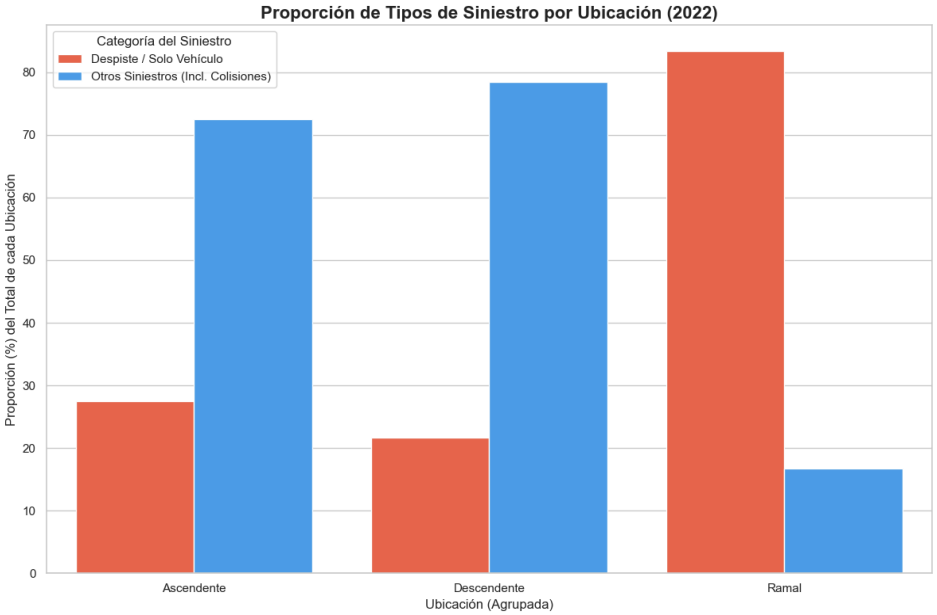
```
--- Test Chi-Cuadrado de Bondad de Ajuste ---
Estadístico Chi-cuadrado (χ²): 29.5849
P-valor: 5.3520098740853275e-08

Resultado (con α=0.05): Se rechaza la Hipótesis Nula (p < 0.05)
La diferencia entre los siniestros observados y los esperados es estadísticamente significativa.
```

Utilizando un nivel de significancia de 0.05, el test arrojó un p-valor inferior a 0.05, motivo por el cual se rechaza la hipótesis nula de igualdad entre las distribuciones observada y esperada. En consecuencia, se concluye que la proporción de siniestros ocurridos en horario pico es significativamente mayor que la que cabría esperar bajo un patrón aleatorio, lo que permite validar la hipótesis planteada.

Hipótesis 2: "El riesgo principal en los ramales no es el choque con otros vehículos, sino el despiste o la colisión de un solo vehículo."

En primer lugar se generó una nueva variable denominada BANDA_GRUPO, que agrupa los valores originales de BANDA Y/O RAMAL en tres categorías operativas: Ascendente, Descendente y Ramal.
Posteriormente, se clasificó el tipo de siniestro en dos grupos analíticos “Despiste / Solo Vehículo” (que incluye las categorías COLISIÓN CON OBSTÁCULO FIJO y SINIESTRO DE UN SOLO VEHÍCULO / SIN COLISION) y “Otros Siniestros (Incl. Colisiones)” (que agrupa los tipos de siniestros restantes)
A partir de estas dos variables categóricas, se construyó una tabla de contingencia y se elaboró una visualización comparativa de proporciones con el objetivo de explorar

descriptivamente la distribución de los tipos de siniestro dentro de cada categoría de ubicación.



Dicho análisis preliminar sugiere una concentración notablemente mayor de siniestros del tipo Despiste / Solo Vehículo en los ramales, lo cual es coherente con la hipótesis planteada.

Con el fin de determinar si esta asociación observada podía atribuirse al azar, se aplicó un test de Chi-cuadrado de independencia.

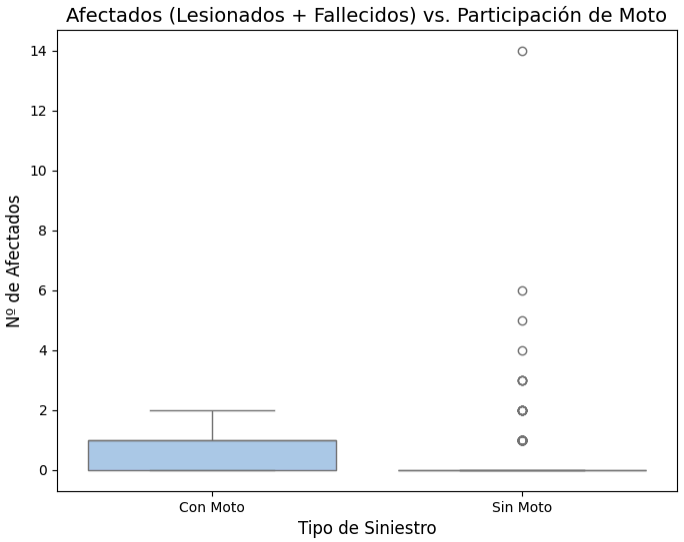
```
--- Resultados del Test ---
Estadístico Chi-cuadrado (χ²): 32.7061
Grados de libertad (dof): 2
Valor p (p-value): 7.906122900927879e-08
```

El test se realizó con un nivel de significancia de α = 0.05. El resultado arrojó un estadístico Chi-cuadrado elevado y un p-valor menor a 0.00001, indicando una evidencia estadística contundente para rechazar la hipótesis nula.

Se concluye que existe una asociación significativa entre la ubicación y el tipo de siniestro, y que en los ramales se registra una proporción considerablemente mayor de siniestros de tipo Despiste / Solo Vehículo. Por lo tanto, los resultados validan la hipótesis planteada, confirmando que este tipo de incidentes constituye el riesgo predominante en los ramales.

Hipótesis 3: “los siniestros que involucren motos tendrán más afectados que los que no”

Se realizó un análisis comparativo de la gravedad de los siniestros, medida por la variable AFECTADOS (suma de lesionados y fallecidos), segmentando los datos en dos grupos: "Con Moto" (siniestros que involucran al menos una motocicleta) y "Sin Moto".



El análisis descriptivo arrojó diferencias notables. El grupo "Sin Moto" presenta una mediana de 0.0 afectados, y dado que tanto el primer cuartil (Q1) como el tercer cuartil (Q3) son también 0.0 1, se infiere que al menos el 75% de los siniestros en esta categoría no tuvieron afectados. La media (0.31) es superior, indicando la presencia de valores atípicos que elevan el promedio, aunque no representan el comportamiento típico del grupo.
En contraste, el grupo "Con Moto" muestra una gravedad significativamente mayor, con una mediana de 1.0 afectado. Esto significa que al menos la mitad de todos los siniestros con moto involucrada resultaron en, como mínimo, una persona afectada. La media (0.76) es más del doble que la del grupo "Sin Moto" y se alinea estrechamente con la mediana.
Para validar estadísticamente esta diferencia, se aplicó un Test U de Mann-Whitney, una prueba no paramétrica adecuada para los datos no normales observados. Se plantearon una Hipótesis Nula de no diferencia y una Hipótesis Alternativa que postula que la distribución de afectados en el grupo "Con Moto" es sistemáticamente mayor.

```
--- Resultados del Test U de Mann-Whitney ---
Estadístico U: 38721.00
P-Valor (p-value): 8.008904849448282e-27

Conclusión: El p-valor es MENOR que 0.05.
Se rechaza la Hipótesis Nula. La diferencia es estadísticamente significativa.
```

El test arrojó un p-valor extremadamente pequeño y muy por debajo del nivel de significancia utilizado 0.05. Por lo tanto, se rechaza la Hipótesis Nula. La diferencia observada no es producto del azar y se concluye, con alta confianza estadística, que los siniestros viales que involucran al menos una motocicleta son significativamente más graves y tienen una tendencia sistemáticamente mayor a producir afectados que aquellos siniestros donde no participan motos.

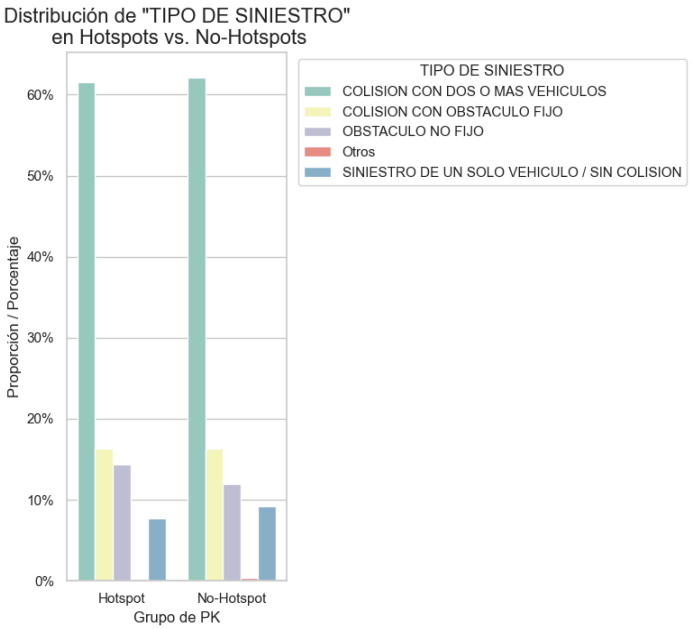
Hipótesis 4: “El tipo de siniestro cambia según si el punto es hotspot o no.”

El presente análisis tuvo como objetivo validar la hipótesis de que los puntos kilométricos catalogados como "hotspots" (definidos como el 10% de los PKs con mayor frecuencia de siniestros) presentan una distribución de TIPO DE SINIESTRO estadísticamente diferente a la de los puntos "no-hotspot". Esta hipótesis busca identificar si "algo" particular en esos lugares (reflejado en el tipo de colisión) que explique su alta siniestralidad.

Se elaboró una tabla de contingencia que resume la relación entre ambas variables. En ella, los tipos de siniestro ocupan las filas y los grupos “Hotspot” y “No-Hotspot” ocupan las columnas. Cada celda representa el número de siniestros de un tipo específico registrados en cada grupo.

```
--- Tabla de Contingencia (Recuentos Observados) ---
GRUPO_HOTSPOT                          Hotspot  No-Hotspot
CARACTERISTICA_AGRUPADA
COLISION CON DOS O MAS VEHICULOS           128         156
COLISION CON OBSTACULO FIJO                 34          41
OBSTACULO NO FIJO                           30          30
Otros                                        0           1
SINIESTRO DE UN SOLO VEHICULO / SIN COLISION 16         23
```

Para complementar esta etapa descriptiva, se generó un gráfico de barras apiladas, que representa visualmente la proporción de cada tipo de siniestro dentro de cada grupo. Para esta visualización, las frecuencias de la tabla de contingencia fueron normalizadas por columna, obteniendo porcentajes en lugar de valores absolutos. Esto permite comparar ambas distribuciones sin que influya la diferencia en la cantidad total de siniestros entre hotspots y no-hotspots.



Para verificar esta asociación, se empleó un Test Chi-Cuadrado de Independencia, la prueba estadística adecuada para evaluar la relación entre dos variables categóricas: la pertenencia a un grupo (Hotspot / No-Hotspot) y el TIPO DE SINIESTRO. La prueba compara las frecuencias observadas en nuestra tabla de contingencia con las frecuencias que se esperarían si no existiera ninguna relación entre las variables (es decir, si fueran independientes).

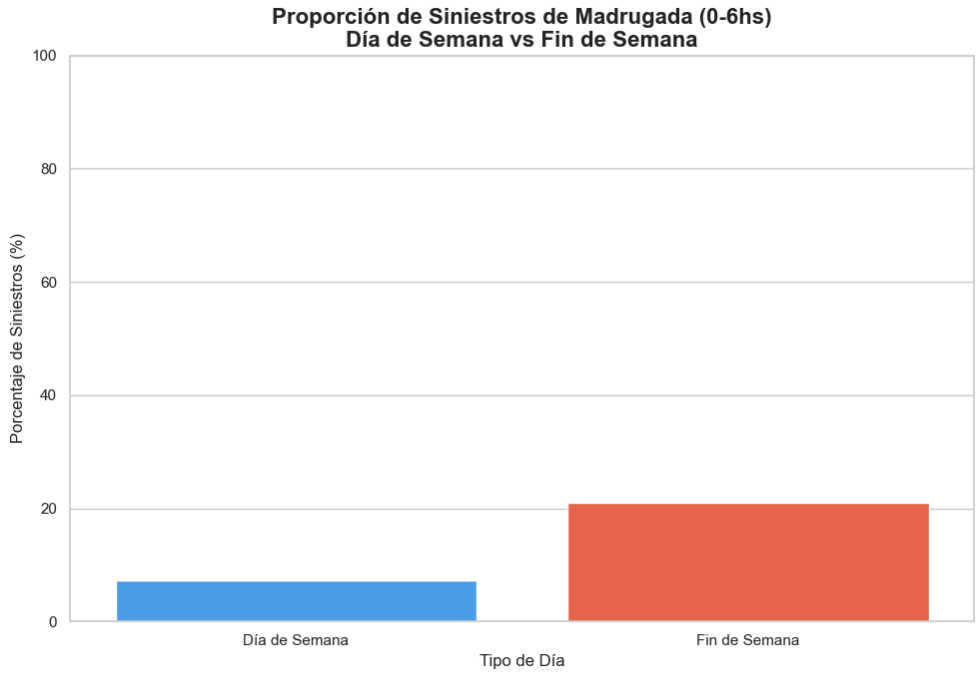
```
--- Resultados del Test Chi-Cuadrado ---
Estadístico Chi-cuadrado (χ²): 1.66
Grados de libertad (dof): 4
P-Valor: 0.7986
...
Se ACEPTA la hipótesis nula.
Conclusión: No se encontró una asociación significativa entre ser un 'hotspot' y el 'TIPO DE SINIESTRO'.
```

Los resultados del test arrojaron que el p-valor es significativamente superior al nivel de significancia (0.05). Por consiguiente, se acepta la hipótesis nula y se concluye que no existe evidencia estadística suficiente para afirmar que haya una asociación entre ser un "hotspot" y el tipo de siniestro que ocurre en él.

En términos prácticos, este hallazgo indica que la distribución de los tipos de siniestro (ej. "COLISIÓN CON DOS O MÁS VEHÍCULOS", "COLISIÓN CON OBSTÁCULO FIJO", etc.) es homogénea a lo largo de las autopistas, independientemente de si un punto kilométrico tiene una frecuencia alta o baja de accidentes. El factor que define a un "hotspot" no es, por lo tanto, un tipo de riesgo específico reflejado en esta variable, sino una mayor frecuencia general de los mismos tipos de siniestros que ocurren en otros lugares.

Hipótesis 5: “Los fines de semanas hay más proporción de siniestros de madrugada que los días de semana”

Para esta hipótesis, se evaluó si la proporción de siniestros viales ocurridos en la franja horaria de madrugada (0–6 hs) difiere entre días de semana y fines de semana.
Para el análisis descriptivo se realizó un gráfico de barras en el que se compararon proporciones de siniestros en madrugada dentro de cada grupo y no cantidades absolutas para así evitar el sesgo generado por la diferencia en cantidad de dias (5 dias de semana o 2 días de fin de semana), así la comparación es justa y estadísticamente válida.



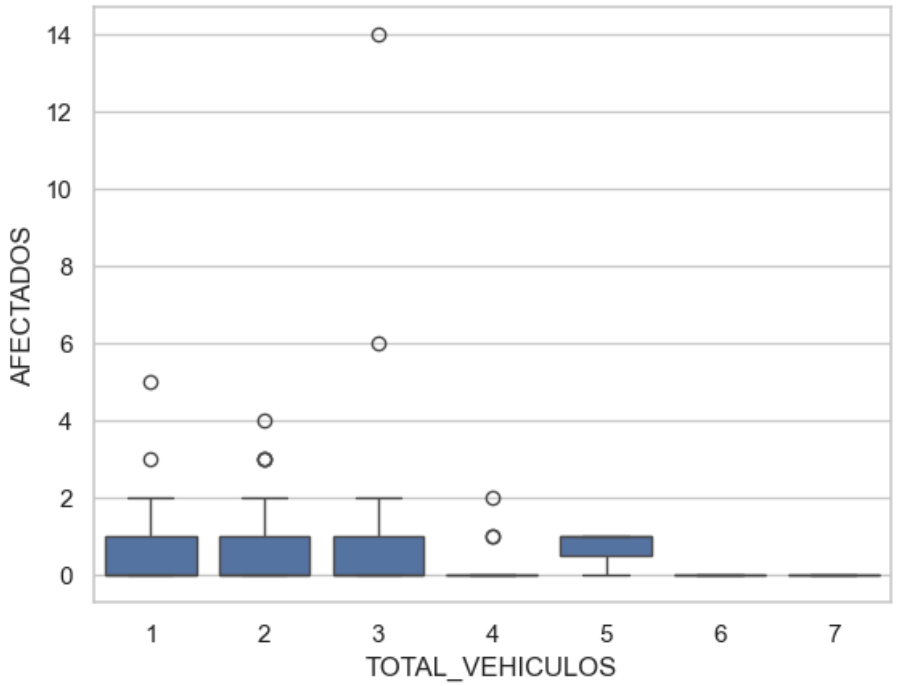
Si bien en el análisis descriptivo se observa una diferencia en las proporciones de siniestros entre los grupos, se realizó un análisis estadístico con el fin de evaluar si dicha diferencia es estadísticamente significativa.

```
Chi-cuadrado (χ²): 15.208
Grados de libertad: 1
p-valor: 9.629288056355261e-05
Resultado: Se RECHAZA la hipótesis nula (H0).
```

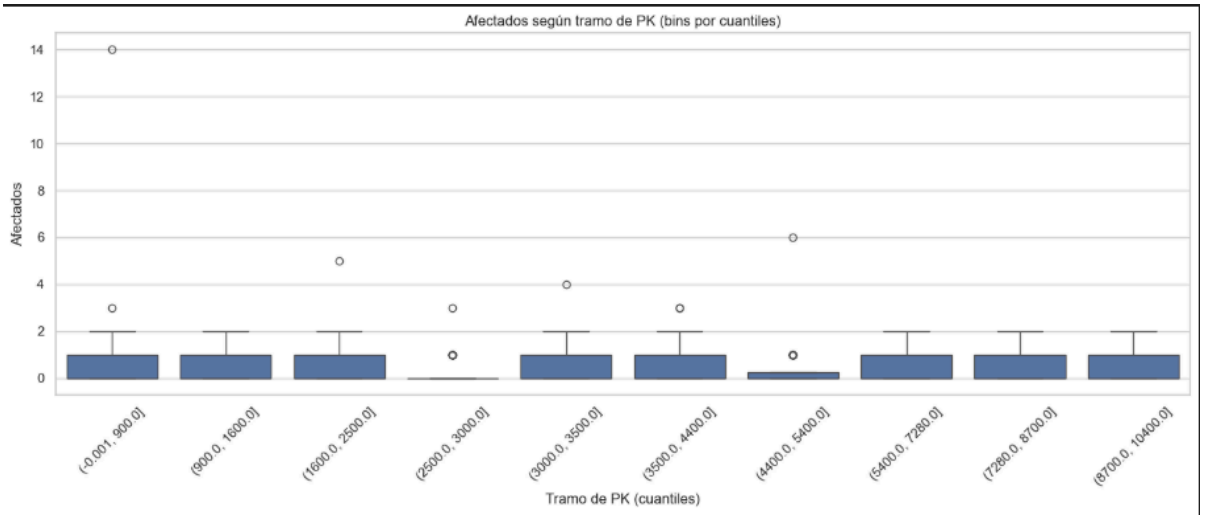
Para ello se construyó una tabla de contingencia entre el tipo de día y la ocurrencia de siniestros en la madrugada, y se aplicó un test Chi-cuadrado de independencia con un nivel de significancia de 0,05. Dado que el p-valor es considerablemente menor que 0,05, se rechaza la hipótesis nula que asume igualdad en la proporción de siniestros de madrugada entre ambos tipos de día. Esto indica que existe evidencia estadísticamente significativa de una asociación entre el tipo de día (semana vs. fin de semana) y la probabilidad de que un siniestro ocurra en la franja de madrugada, sugiriendo que la madrugada del fin de semana presenta un comportamiento diferente al de los días hábiles.

Hipótesis 6: “El número de afectados en un siniestro está vinculado a la cantidad de vehículos involucrados y al punto kilométrico donde ocurrió”

Para investigar si la cantidad de vehículos involucrados en un siniestro y el punto kilométrico (PK) influyen en la probabilidad de que haya personas afectadas, se realizó primero un análisis exploratorio con visualizaciones.

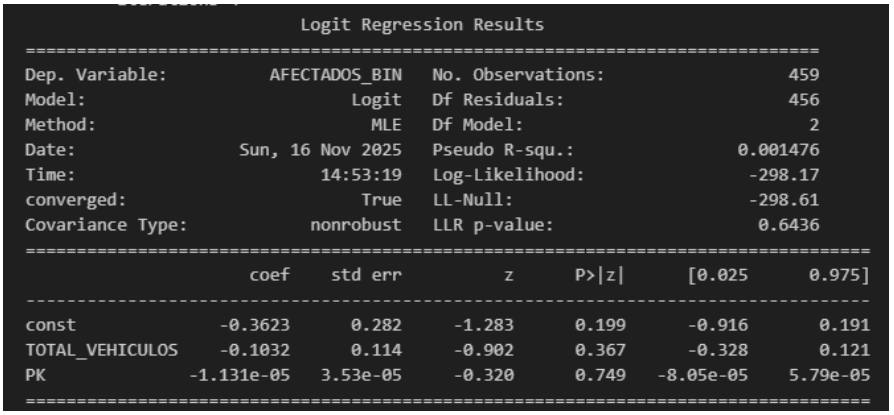


El boxplot que relaciona TOTAL_VEHICULOS con el número de afectados muestra que los grupos más frecuentes (siniestros con 1, 2 o 3 vehículos) presentan prácticamente la misma distribución: la gran mayoría de los casos tienen cero afectados, las medianas se mantienen en el mismo nivel y las diferencias entre grupos aparecen únicamente en algunos valores atípicos. Además, al aumentar el número de vehículos a 4, 5, 6 o 7, no se observa un crecimiento en la cantidad de afectados; incluso con varios vehículos involucrados, la distribución sigue concentrándose en valores muy bajos. En conjunto, el gráfico sugiere que no hay una tendencia clara que indique que involucrar más vehículos aumente la cantidad de lesionados o fallecidos.



Luego se analizó si la localización del siniestro podría tener algún efecto. Para ello, el PK fue agrupado en tramos por cuantiles, logrando bins con tamaños comparables. El boxplot

resultante muestra una distribución prácticamente idéntica de afectados en todos los tramos: la mediana es cero en cada segmento, los valores típicos permanecen bajos y los casos más altos corresponden a eventos puntuales. Esto implica que el PK, por sí solo, tampoco parece estar asociado a una mayor gravedad del siniestro.



Logit Regression Results

Dep. Variable:	AFECTADOS_BIN	No. Observations:	459
Model:	Logit	Df Residuals:	456
Method:	MLE	Df Model:	2
Date:	Sun, 16 Nov 2025	Pseudo R-squ.:	0.001476
Time:	14:53:19	Log-Likelihood:	-298.17
converged:	True	LL-Null:	-298.61
Covariance Type:	nonrobust	LLR p-value:	0.6436

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3623	0.282	-1.283	0.199	-0.916	0.191
TOTAL_VEHICULOS	-0.1032	0.114	-0.902	0.367	-0.328	0.121
PK	-1.131e-05	3.53e-05	-0.320	0.749	-8.05e-05	5.79e-05

Para contrastar formalmente estas observaciones, se realizó un modelo de regresión logística donde la variable dependiente indica si hubo o no personas afectadas, y las variables independientes fueron la cantidad total de vehículos involucrados y el PK. Los resultados estadísticos confirman lo observado visualmente: tanto TOTAL VEHICULOS como PK tienen p valores muy superiores al 0.05 (0,367 y 0,749 respectivamente), lo que indica ausencia de evidencia estadística de relación mostrando que el modelo prácticamente no mejora la predicción respecto a no incluir ninguna variable.
En conjunto, los gráficos exploratorios y el análisis estadístico llevan a la misma conclusión: ni la cantidad de vehículos involucrados ni la ubicación PK permiten explicar o predecir la presencia de afectados en un siniestro. La severidad parece estar determinada por otros factores no contemplados aquí, como el tipo de maniobra, la velocidad o las características del siniestro.

Conclusión

El análisis estadístico realizado permitió identificar patrones consistentes de riesgo asociados a factores temporales, geográficos y vinculados al tipo de vehículo involucrado. En conjunto, los resultados muestran que la ocurrencia y la gravedad de los siniestros no se distribuyen de manera uniforme, sino que responden a condiciones específicas del entorno vial.

En primer lugar, se observó un riesgo temporal marcado: los siniestros se concentran en las horas pico, reforzando la relación entre congestión vehicular y accidentalidad. A su vez, la proporción de accidentes ocurridos en la franja de madrugada (0–6 hs) es significativamente mayor los fines de semana que durante los días hábiles, lo cual sugiere un patrón asociado a la conducción nocturna recreativa, posiblemente influida por fatiga, consumo de alcohol o menor percepción de riesgo.

En cuanto al riesgo geográfico, el análisis por ubicación reveló diferencias notables entre bandas y ramales. Aunque en términos absolutos los ramales concentran menos siniestros, el

tipo de accidente predominante en ellos es el “siniestro de un solo vehículo”, a diferencia de las bandas ascendente y descendente, donde prevalecen las colisiones múltiples. Esto implica que los ramales presentan condiciones geométricas o operativas que favorecen los despistes, por lo que podrían beneficiarse de mejoras en señalización, iluminación o diseño vial.

Respecto a la gravedad de los siniestros, el estudio confirmó que la presencia de motocicletas incrementa de manera significativa el número de personas afectadas, destacando a este tipo de vehículo como un factor crítico en términos de vulnerabilidad. En contraste, se encontró que ni el punto kilométrico (PK) ni el número total de vehículos involucrados permiten explicar o predecir adecuadamente la ocurrencia de lesionados o fallecidos, tal como lo evidenció el modelo de regresión logística, cuyos coeficientes no resultaron estadísticamente significativos.

Finalmente, el análisis de hotspots mostró que los puntos críticos de accidentalidad se definen principalmente por su mayor frecuencia de siniestros, y no por presentar un tipo de riesgo cualitativamente distinto al del resto de lugares. Esto sugiere que la concentración de accidentes responde más al volumen de exposición que a un patrón específico de causalidad.

En conjunto, estos hallazgos ofrecen una base sólida para orientar futuras intervenciones de seguridad vial. Para AUSA, las prioridades deberían focalizarse en: la gestión del flujo vehicular en horas pico, campañas específicas para la conducción nocturna de fin de semana, intervenciones estructurales en ramales para reducir despistes y estrategias de protección dirigidas a motociclistas, el grupo de usuarios con mayor vulnerabilidad según los datos analizados.