Proyecto final - Lenguajes Minería de Datos

Colegio Universitario de Cartago

Profesor: Osvaldo Gonzalez Chaves

Javier Brenes Redondo

Diego Montero Vargas

Primer Cuatrimestre, 2025

Contenido

Introducción

Este proyecto se centra en la aplicación de las técnicas aprendidas en este curso para analizar el conjunto de datos "ARCHIVO_VIOLENCIA_DOMESTICA.tsv". Utilizando bibliotecas de R, el objetivo principal es extraer información relevante y responder a las preguntas de investigación planteadas, además de generar estadísticas descriptivas que permitan comprender mejor los datos.

Conjunto de datos

El archivo seleccionado, "ARCHIVO_VIOLENCIA_DOMESTICA.tsv", contiene registros de denuncias por violencia doméstica a nivel nacional desde 2015 hasta 2025, obtenidos del portal del Organismo de Investigación Judicial. A continuación, se explican las variables presentes en este conjunto de datos.

Variable	Tipo	Descripcion
Anno	int	Año de la estadística
Mes	int	Mes de la estadística
NombreMes	texto	Descripción del mes
NombreMateria	texto	Nombre de la materia
NombreCircuito	texto	Nombre del circuito
NombreDespacho	texto	Descripción del despacho
NombreTipoDespacho	texto	Descripción del tipo de despacho
CirculanteInicial	int	Cantidad de expedientes activos al iniciar el mes
CirculanteInicialLeg	int	Subconjunto del apartado anterior, en donde el tipos de caso es "Legajo", válido para despachos laborales y contenciosos.
Entrados	int	Cantidad de expedientes entrados como nuevos durante el mes.

TestimoniosPiezas	int	Subconjunto de los entrados, en donde el expediente sea un testimonio de piezas.
Legajos	int	Subconjunto de los entrados, en donde el expediente sea un "legajo".
Reentrados	int	Cantidad de expedientes reentrados durante el mes.
TerminadosXImcompetencia	int	Subconjunto de los terminados, en donde el motivo de término sea "Por Incompetencia"
AbandonadosOInactivos	int	Subconjunto de los terminados, en donde el motivo de término sea "Abandonado ó Inactivo"
Terminasdos	int	Cantidad de expedientes finalizados durante el mes.
CirculanteFinal	int	Cantidad de expedientes activos al finalizar el mes.
CirculanteFinalLegajos	int	Subconjunto del apartado anterior, en donde el tipos de caso es "Legajo", válido para despachos laborales y contenciosos.

Estadísticas Básicas

Para aquellos valores numéricos del dataset se encuentran las siguientes estadísticas

variable	Media	Mediana	Moda
CirculanteInicial	587.7595905	386	84
CirculanteInicialLeg	0.0000000	0	0
Entrados	61.6674479	40	9
TestimoniosPiezas	13.0511903	3	0
Legajos	0.0000000	0	0
Reentrados	0.5901073	0	0
TerminadosXImcompetencia	8.2380659	1	0
AbandonadosOInactivos	0.0000000	0	0
Terminasdos	67.5211546	38	0
CirculanteFinal	587.1793512	386	43
CirculanteFinalLegajos	0.0000000	0	0

Correlaciones

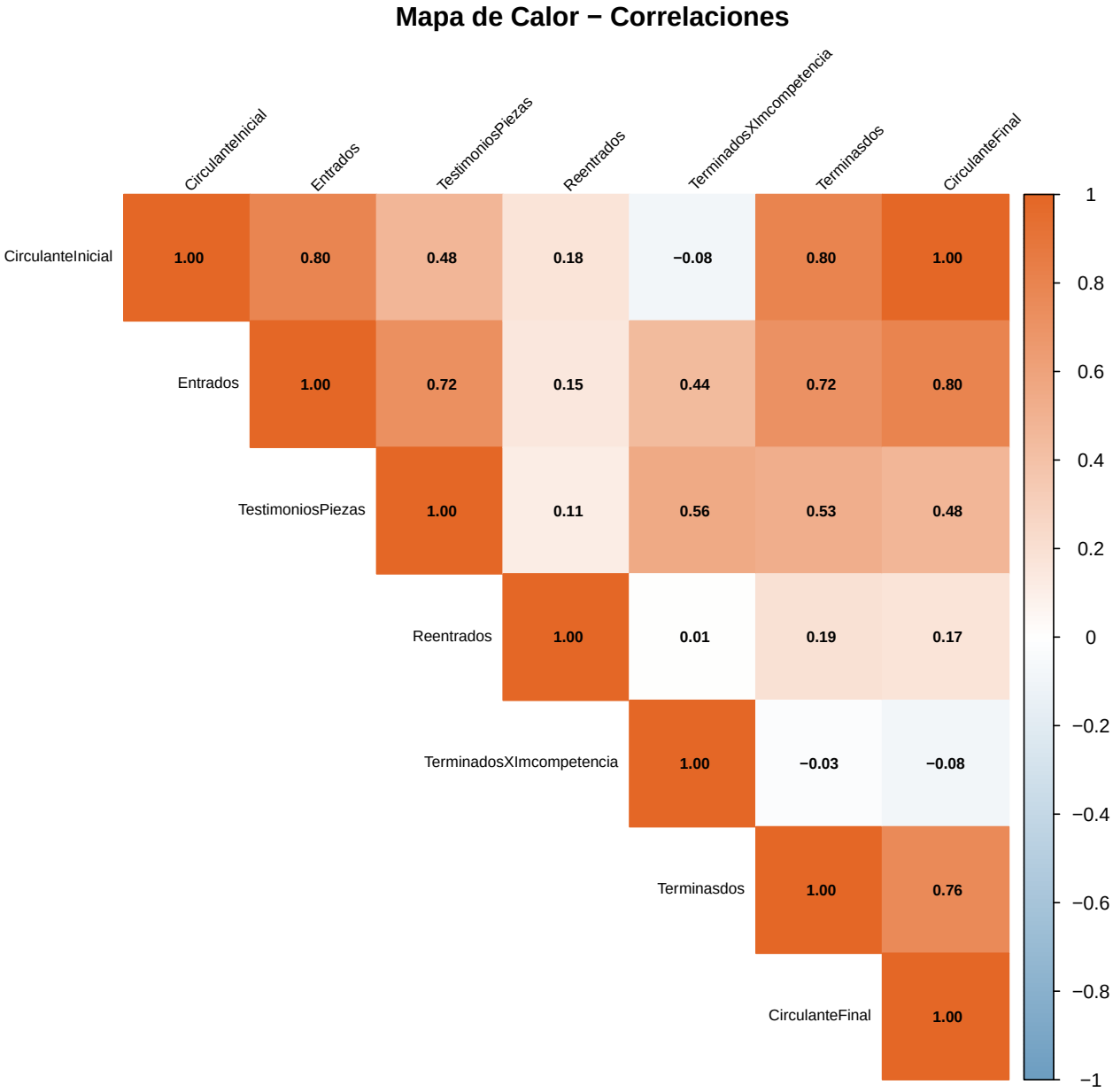
Para analizar las relaciones entre las variables numéricas del conjunto de datos, se presenta la siguiente matriz de correlaciones:

	CI	E	TP	R	TxI	T	CF
CI	1.0000000	0.7998813	0.4776897	0.1753781	-0.0801543	0.8010426	0.9972758
E	0.7998813	1.0000000	0.7217604	0.1542879	0.4419779	0.7167101	0.8010366
TP	0.4776897	0.7217604	1.0000000	0.1105398	0.5565387	0.5271461	0.4774376
R	0.1753781	0.1542879	0.1105398	1.0000000	0.0093771	0.1901260	0.1713953

	CI	E	TP	R	TxI	T	CF
TxI	-0.0801543	0.4419779	0.5565387	0.0093771	1.0000000	-0.0292947	-0.0806422
T	0.8010426	0.7167101	0.5271461	0.1901260	-0.0292947	1.0000000	0.7612602
CF	0.9972758	0.8010366	0.4774376	0.1713953	-0.0806422	0.7612602	1.0000000

Diccionario:

- CI: CirculanteInicial
- E: Entrados
- TP: TestimoniosPiezas
- R: Reentrados
- TxI: TerminadosXImcompetencia
- T: Terminasdos
- CF: CirculanteFinal



Interpretación de las correlaciones:

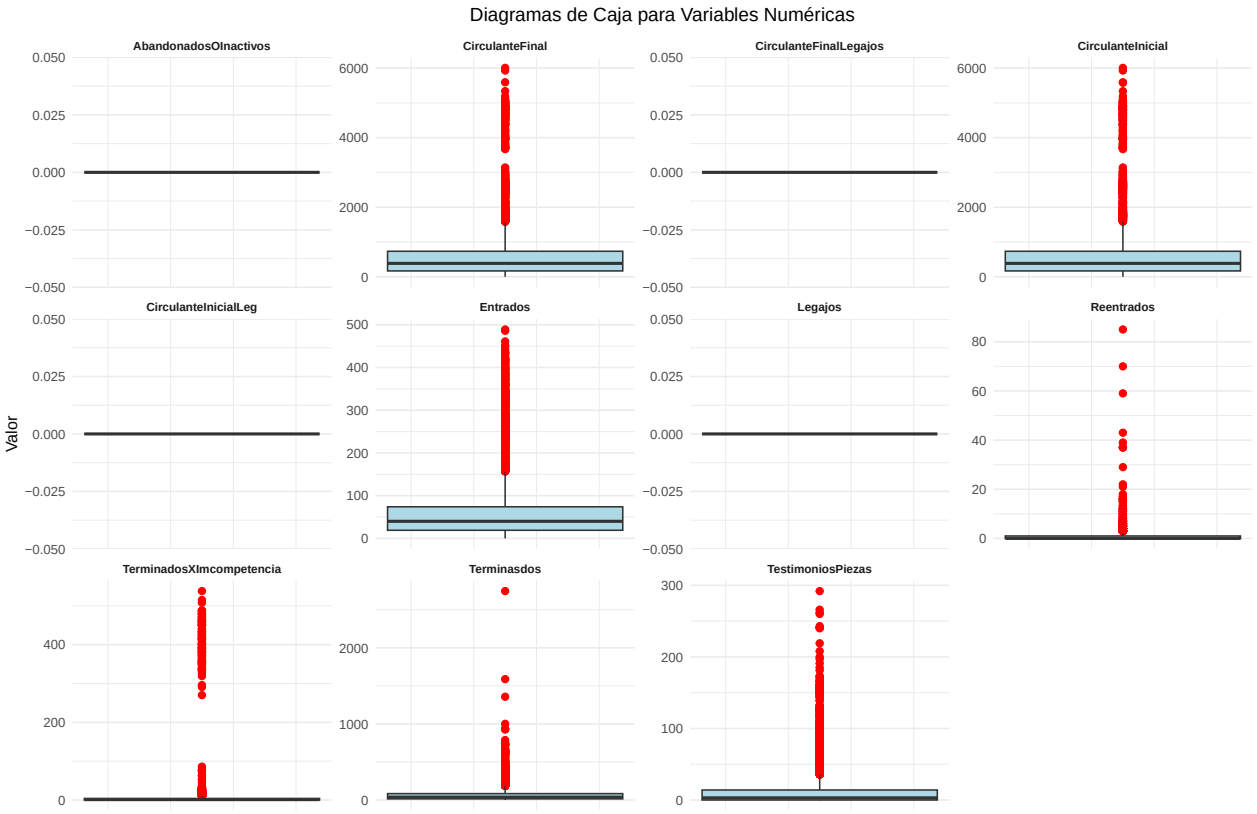
- Los valores cercanos a 1 (naranja oscuro) indican una correlación positiva fuerte
- Los valores cercanos a -1 (azul oscuro) indican una correlación negativa fuerte
- Los valores cercanos a 0 (blanco) indican poca o ninguna correlación
- El tamaño y color de los círculos representan la fuerza y dirección de la correlación

Como se observa en el gráfico en su mayoría las variables tienen relación entre ellas de una forma positiva, tienen a crecer juntas.

Nota: Para el cálculo de las correlaciones, se eliminaron las columnas `CirculanteInicialLeg`, `Legajos`, `AbandonadosOInactivos` y `CirculanteFinalLegajos` del conjunto de datos. Estas columnas contenían únicamente el valor 0, lo que significa que no presentaban ninguna variación. Como la variación es fundamental para calcular la correlación (se utiliza en una división), una columna sin variación causaría un error matemático (división por cero). Por esta razón, se optó por remover estas columnas para evitar problemas en el análisis.

Valores atípicos

Para visualizar la distribución y detectar valores atípicos en las variables numéricas, se presentan los siguientes diagramas de caja:



Los diagramas de caja nos permiten observar: - La distribución de los datos para cada variable - La presencia de valores atípicos (puntos rojos) - La mediana y los cuartiles de cada variable

Hipotesis

Hipótesis 1

Planteamiento Distribución Geográfica de la Violencia Doméstica:

Pregunta de Investigación: ¿Existe una variación significativa en la incidencia de casos de violencia doméstica entre áreas internas del país (San Jose, Heredia, Alajuela, Cartago) y áreas externas del país (Puntarenas, Limon, Guanacaste)?

Objetivos: Analizar la distribución geográfica de los casos nuevos y terminados de violencia doméstica por circuito judicial. Realizar una comparación de la incidencia de casos entre áreas internas del país y áreas externas.

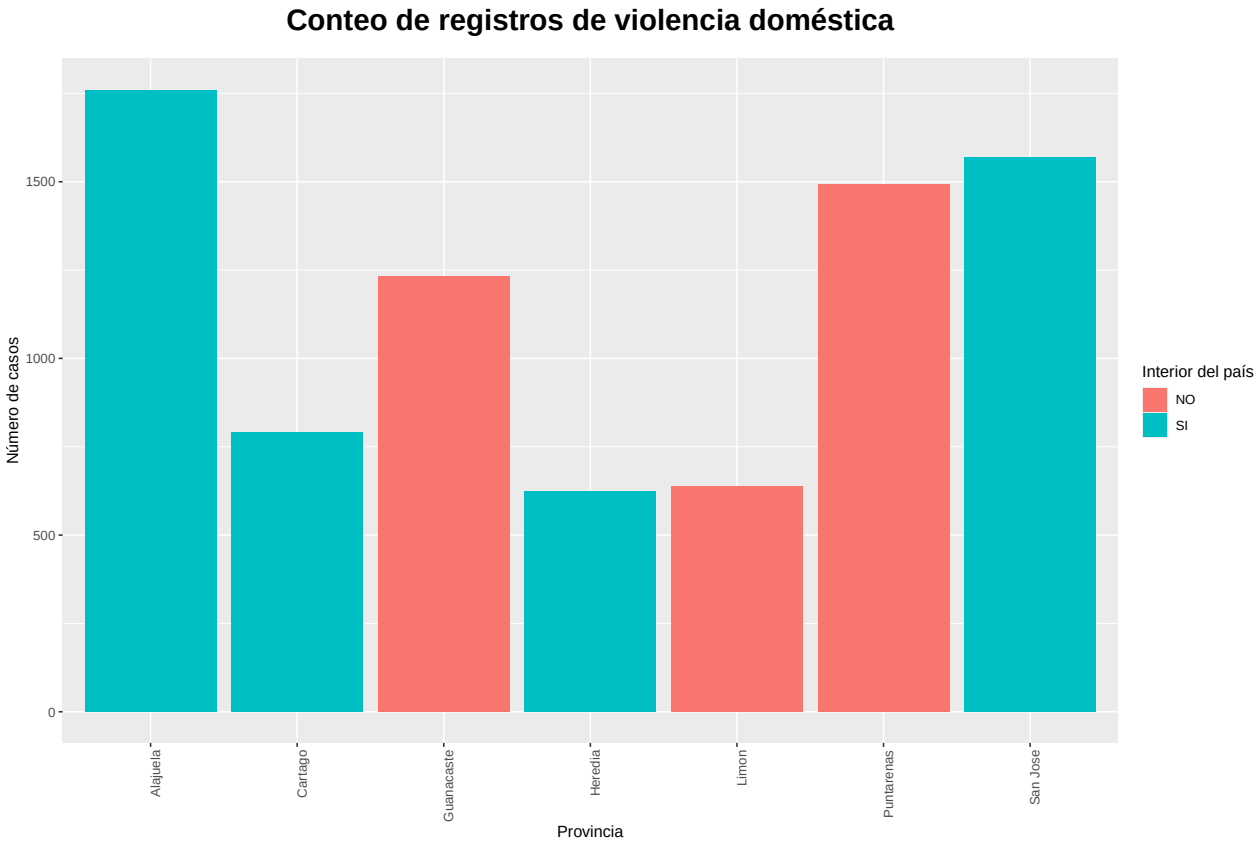
Requerimientos de Datos: Depuración y homogeneización del conjunto de datos existente. Enriquecimiento del conjunto de datos para permitir la clasificación de los circuitos judiciales por provincia.

Análisis Nuestro dataset no incluye una columna de provincia, es por eso que tuvimos que enriquecer el dataset a través de analizar `NombreCircuito` y apartir de este determinar la provincia en la que se encontraba así como clasificar si dicha provincia se encuentra en el interior del país o en el exterior:

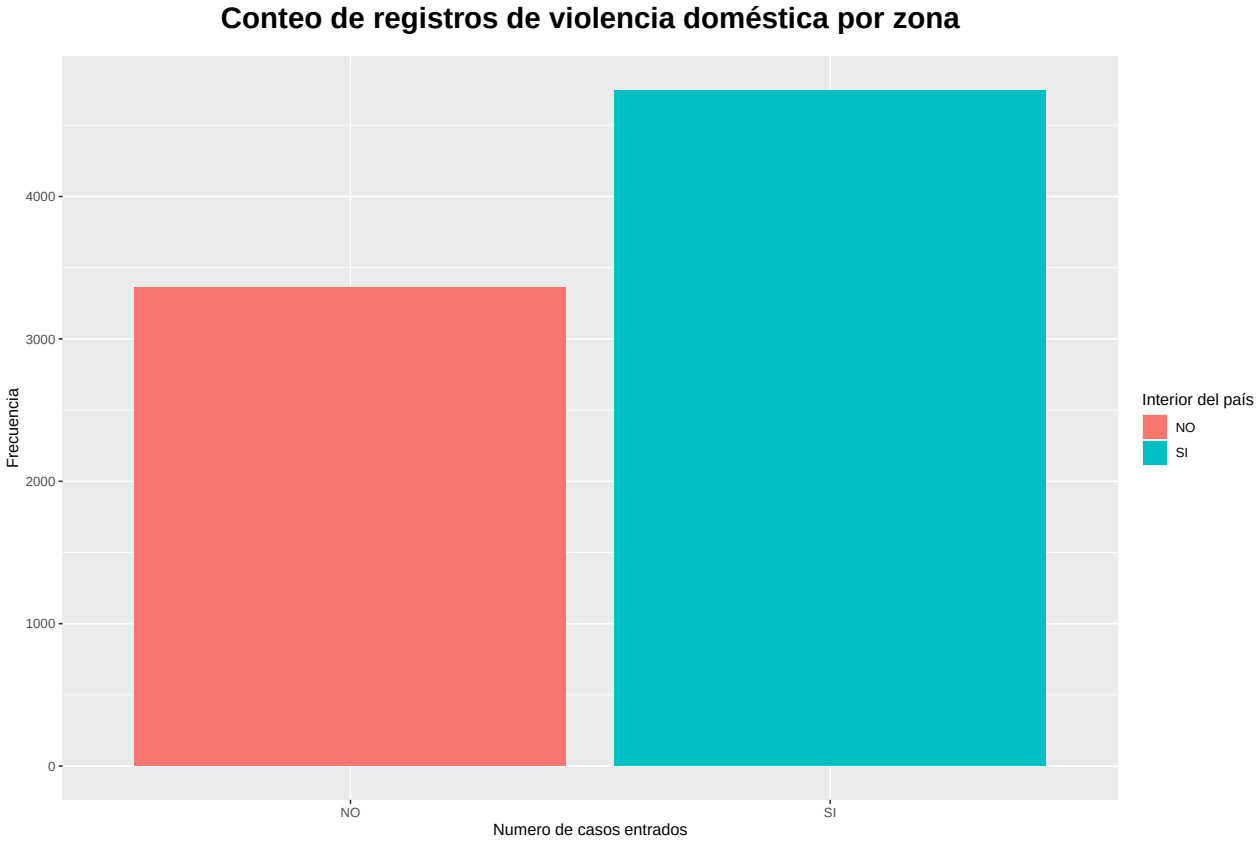
Anno	Mes	NombreCircuito	Entrados	Terminasdos	Provincia	InteriorPais
2015	1	CIRCUITO JUDICIAL DE GOLFITO	99	82	Puntarenas	NO
2015	1	CIRCUITO JUDICIAL DE ATENAS	27	12	Alajuela	SI
2015	1	II CIRCUITO JUDICIAL DE GUANACASTE (NICOYA)	90	76	Guanacaste	NO
2015	1	I CIRCUITO JUDICIAL DE LA ZONA SUR (PEREZ ZELEDON)	123	99	San Jose	SI
2015	1	II CIRCUITO JUDICIAL DE ALAJUELA (SAN CARLOS)	92	185	Alajuela	SI
2015	1	II CIRCUITO DE LA ZONA ATLÁNTICA (POCOCÍ-SIQUIRRES)	126	146	Limon	NO
2015	1	I CIRCUITO JUDICIAL DE GUANACASTE (LIBERIA)	88	84	Guanacaste	NO
2015	1	CIRCUITO JUDICIAL DE TURRIALBA	69	160	Cartago	SI
2015	1	CIRCUITO JUDICIAL DE GOLFITO (PUERTO JIMENEZ)	28	29	Puntarenas	NO
2015	1	CIRCUITO JUDICIAL DE HEREDIA (SARAPIQUÍ)	50	111	Heredia	SI

Nota: Primeros 10 registros del dataset con las nuevas columnas.

Iniciemos haciendo un conteo por provincia

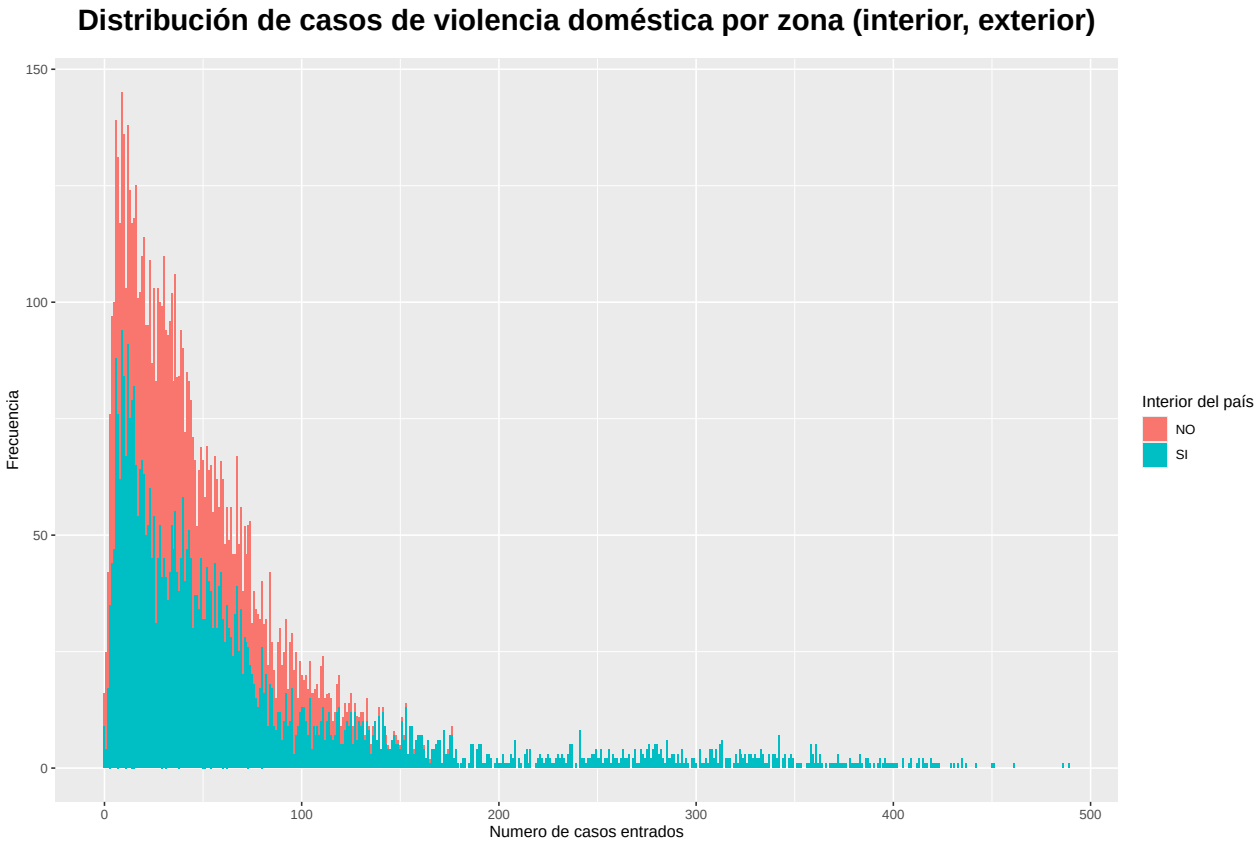


Luego un conteo por zona



Si nos basaramos en un conteo de registros, es claro que en el interior del país hay mas casos de violencia domestica, pero esto no seria una interpretacion valida, ya que en el interior del país estamos contando mas provincias que en el exterior del país, ademas cabe la posibilidad que en el interior hayan mas circuitos por lo que es normal que hayan mas registros en dataset.

Para poder responder la pregunta que se planteo necesitamos analizar la variable `Entrados` y su distribución por zona.

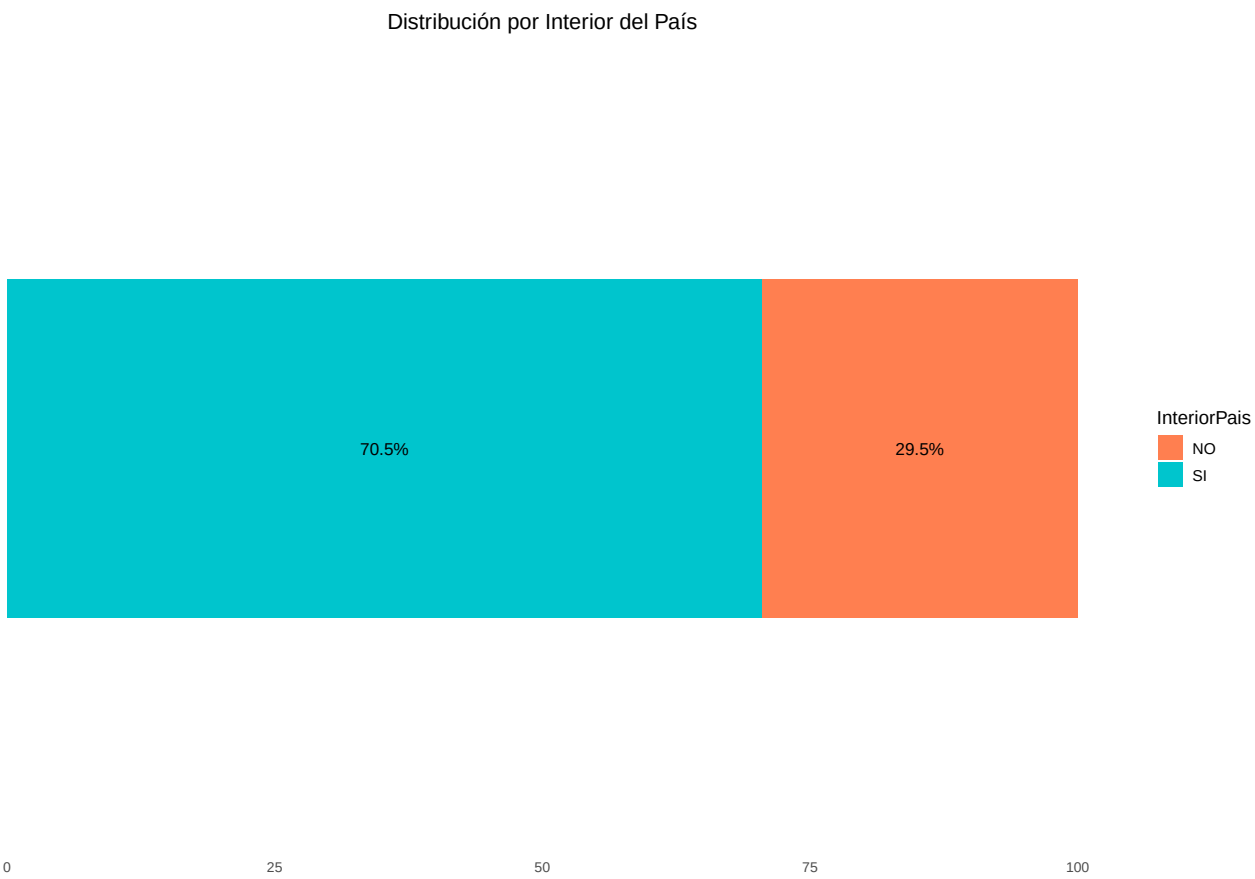


Como se aprecia en el gráfico anterior, en el exterior del país (Guanacaste, Puntarenas, Limón) es donde mas casos entrados de violencia domestica hay en una mayor frecuencia pero en rangos menores, en comparacion al interior del país donde la cantidad de casos entrados en rangos mayores entre 160 y 500 predomina.

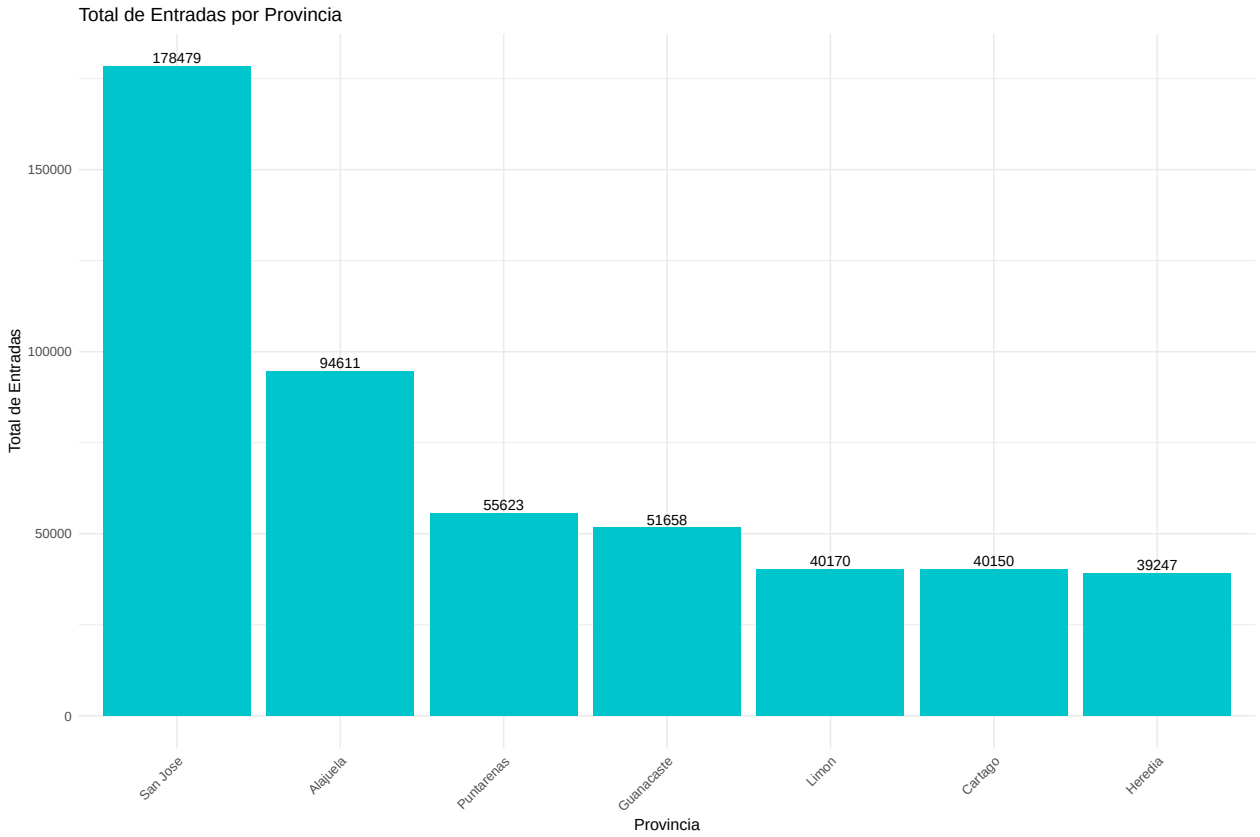
Veamos un histograma para ver si existe un crecimiento a lo largo de los años.

Si vemos los datos desde un punto de vista historico, se aprecia que internamente es donde mas casos de violencia domestica existen, tambien que año a año los valores incrementan.

Pero si hacemos una suma de todos los casos entrantes a lo largo de los años, agrupado por zona, que obtenemos ?



Es posible conocer la totalidad de casos por Provincia entonces ?



Conclusión Podemos definir que la mayoría de los casos de violencia domestica se presentan en el interior del país, donde luego de nuestro análisis San Jose es la provincia con mas casos registrados.

Aunque los datos muestran que en el exterior del país hay en frecuencia mas denuncias, se ve opacado por la cantidad de casos en rangos altos que se da en el interior del país.

Hipótesis 2

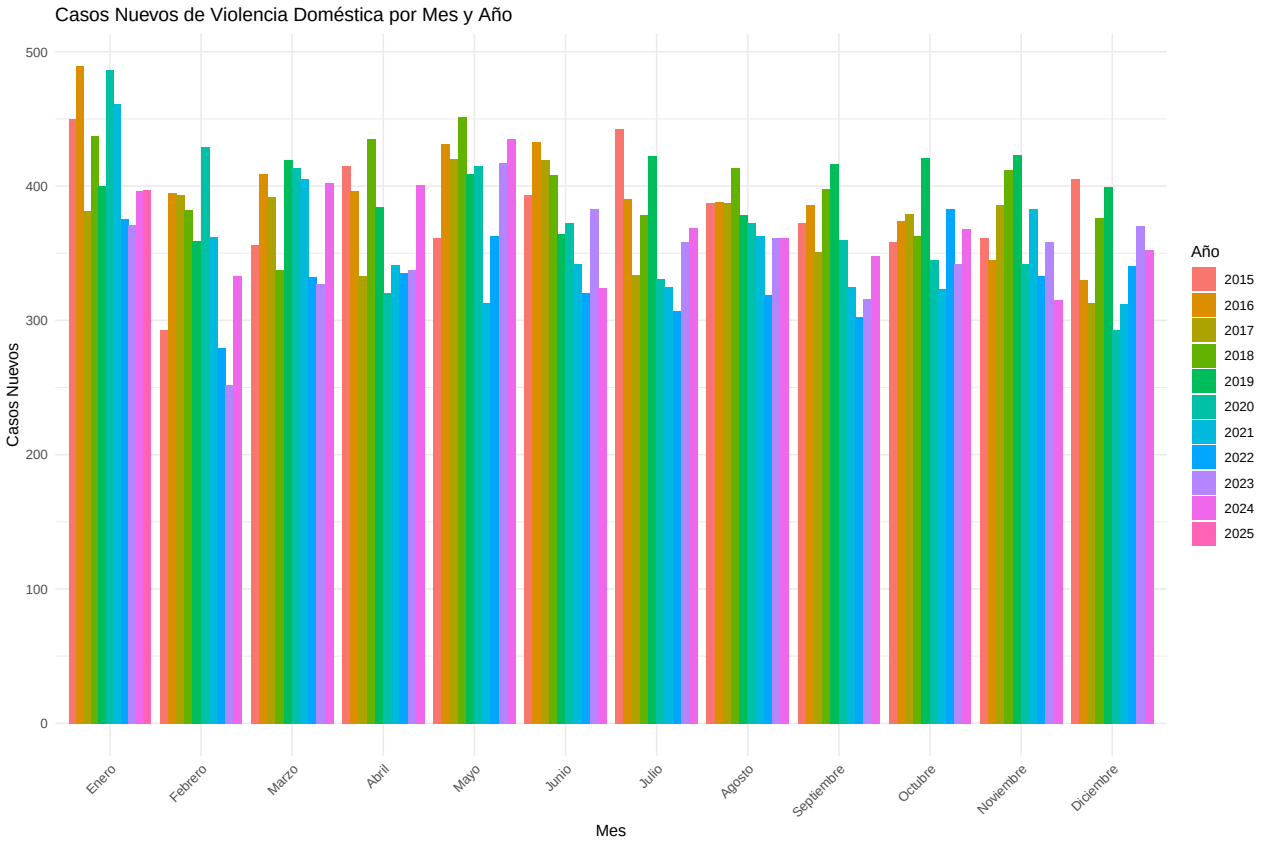
Planteamiento Temporalidad de la Violencia Doméstica:

Pregunta de Investigación: ¿Se identifican periodos específicos del año con un aumento significativo en la presentación de casos de violencia doméstica?

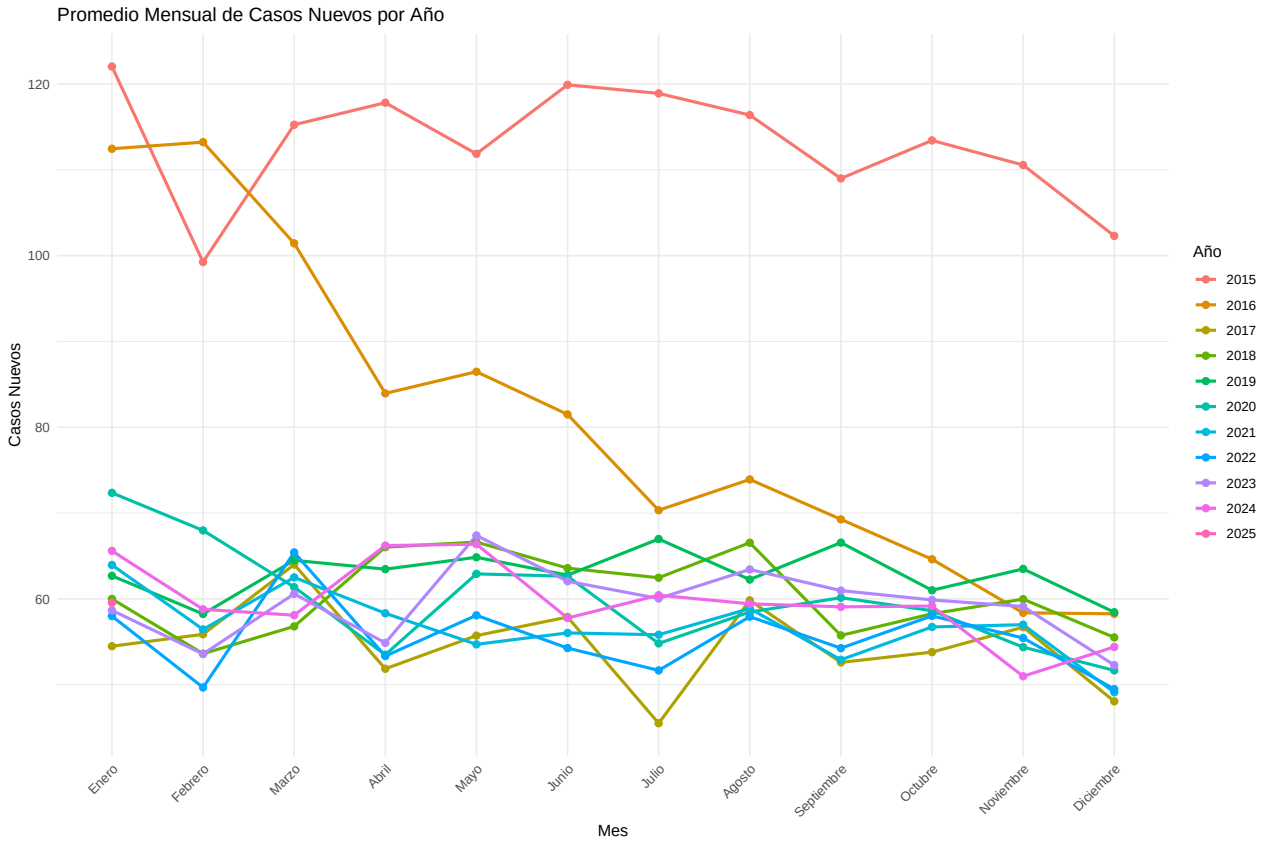
Objetivos: Determinar los periodos del año (meses) con mayor incidencia de casos nuevos de violencia doméstica. Analizar la evolución de los casos nuevos a lo largo de los años.

Requerimientos de Datos: Definición de un subconjunto de datos a partir del existente, ordenado cronológicamente por mes y año de ingreso de los casos.

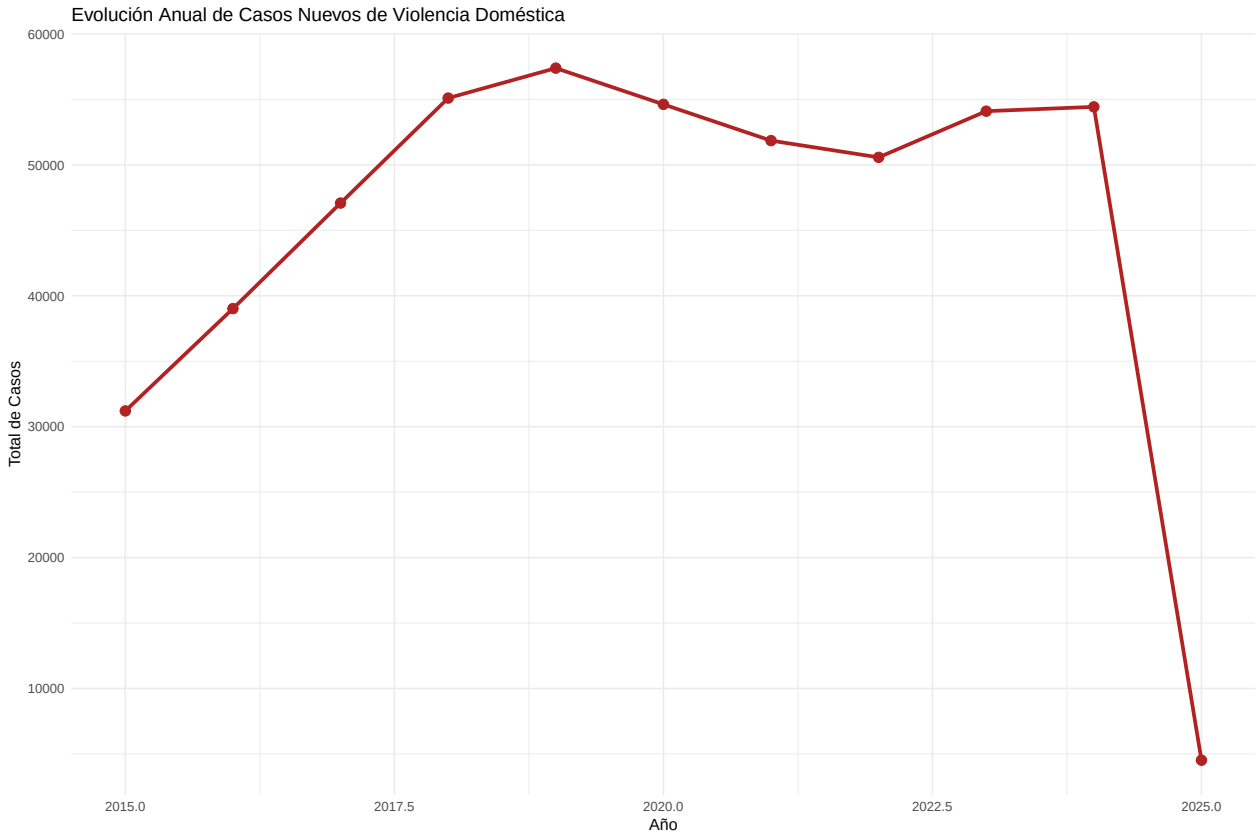
Análisis Este gráfico muestra un conteo mensual desde el año 2015 al 2025 con el total de nuevos casos de violencia Es importante destacar que los datos proporcionados del 2025 llegan solamente del mes de Enero



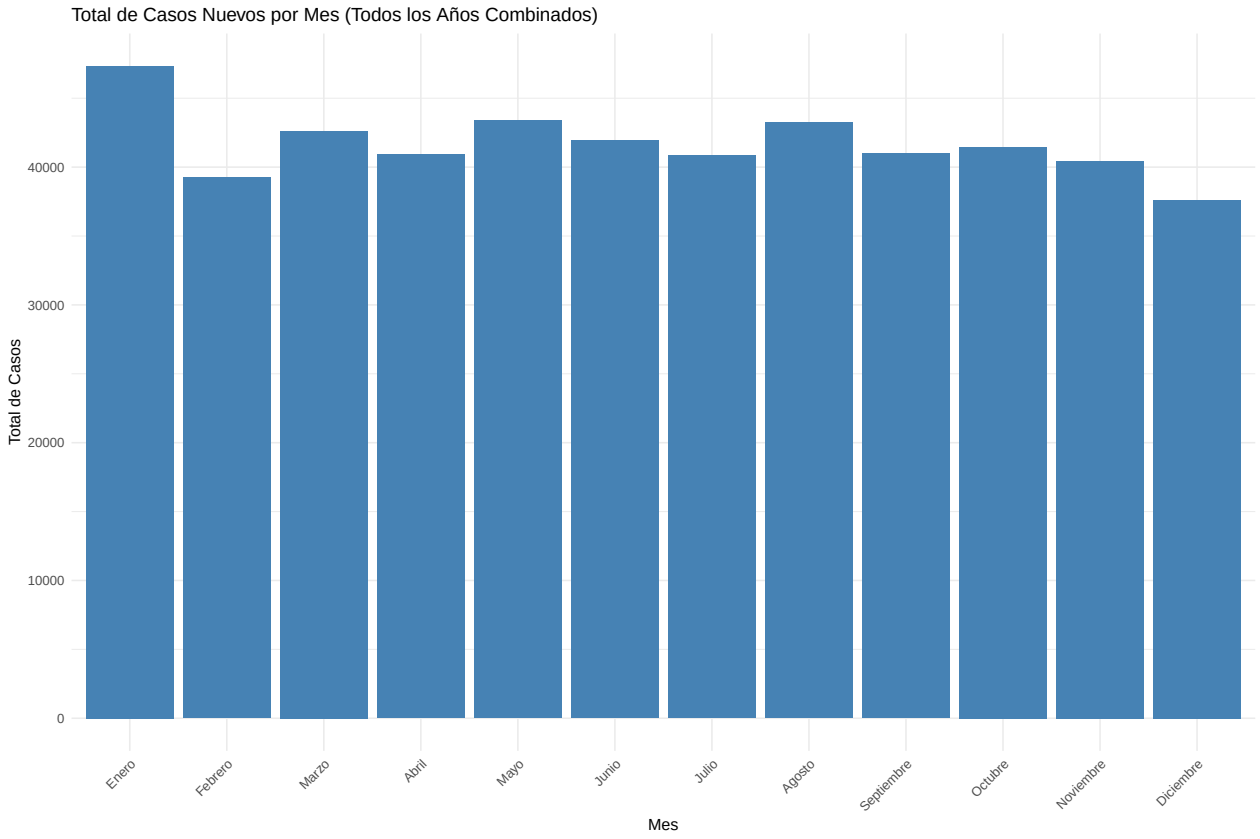
En este gráfico podemos ver el promedio mensual de nuevos casos por cada año. Encontramos que el 2015 y 2016 sobresalen del promedio de los otros años



En este gráfico se contabilizan todos los casos nuevos de cada año Se encuentra que desde el 2015 hasta 2019 hay una tendencia a la alza de casos y luego cae levemente



Este gráfico muestra el total de casos nuevos separados por mes



Conclusión Se confirma que en los últimos 10 años el mes con mas casos de denuncias por violencia doméstica es el mes de Enero en primer lugar, seguido de Mayo y Agosto y los demás meses se mantienen muy similares entre ellos. Mientras que los dos mas bajos son Diciembre y Febrero.

Hipótesis 3

Planteamiento Eficiencia en la Resolución de Casos

Pregunta de Investigación: ¿Cuál es la eficiencia de los diferentes circuitos judiciales en la resolución de casos de violencia doméstica en el 2024?

Objetivos: Evaluar la eficiencia de los despachos judiciales en el cierre de casos de violencia doméstica. Identificar los circuitos judiciales con mayor y menor eficiencia en la resolución de casos.

Requerimientos de Datos: Utilización del conjunto de datos existente para analizar la relación entre los casos ingresados y los casos terminados por circuito judicial.

Análisis Como nos lo hemos planteado queremos analizar unicamente el año 2024 es por eso que iniciamos filtrando estos datos para obtener solo lo requerido.

Para poder determinar la eficiencia de cada circuito hemos definido las siguientes formulas a partir de la siguientes variables:

1. `CirculanteInicial`
2. `Entrados`
3. `Terminasdos`
4. `TerminadosXImcompetencia`

Primer Formula Cuantos casos fueron resueltos comparados al total de trabajo (circulantes + entrados), donde un valor alto indica mejor eficiencia en manejo de caso.

```
Eficiencia_Carga_Total = (Total_Terminados / Carga_Total) * 100
```

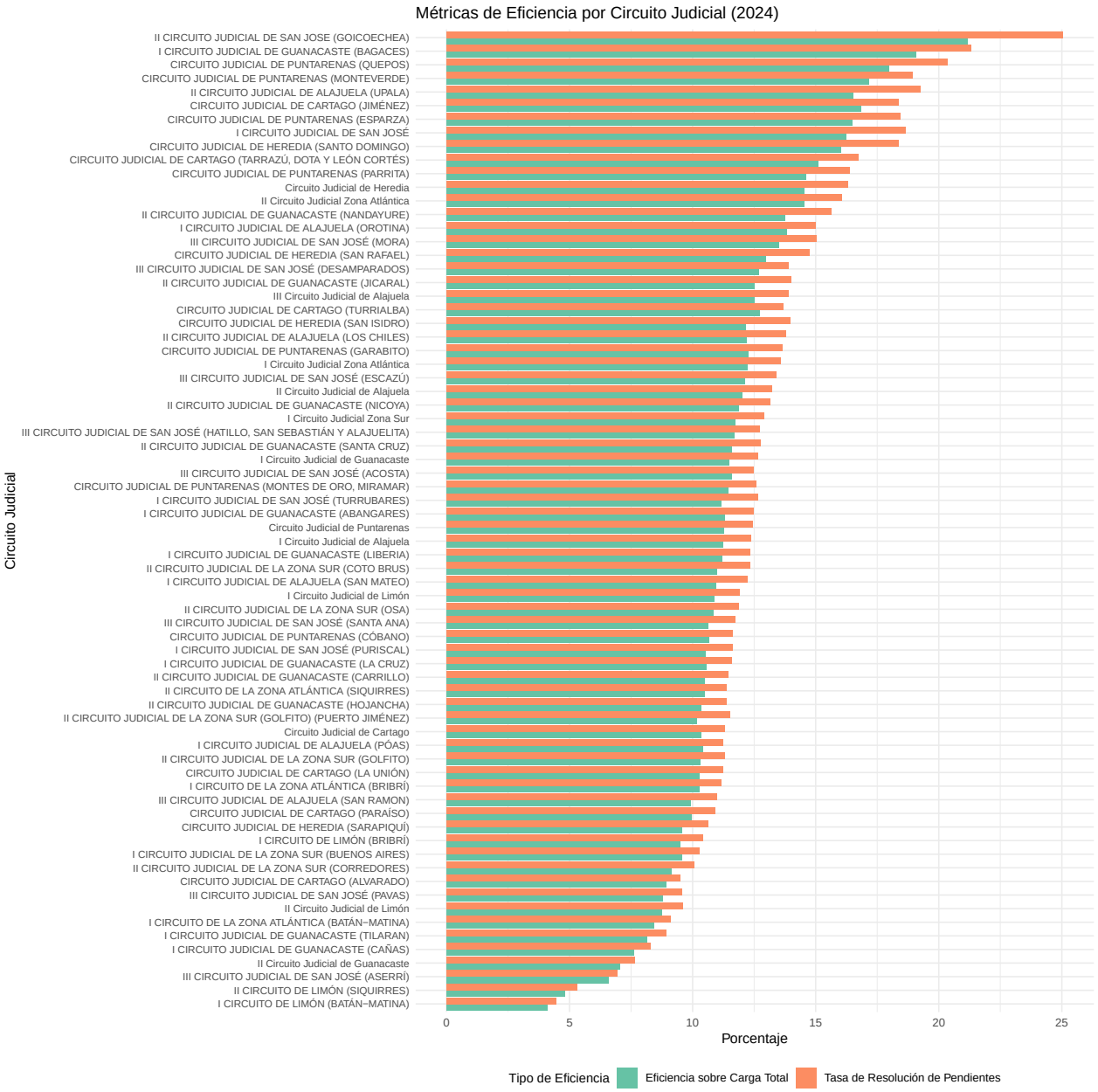
Segunda Formula Que tanto manejan la cantidad de casos, mejor porcentaje indica mejor forma de manejar casos pendientes

```
Tasa_Resolucion_Pendientes = (Total_Terminados / Total_Circulante_Inicial) * 100
```

Lo primero que haremos es sumarizar los datos

Una vez que tenemos los datos sumarizados queremos convertir las columnas en filas, esto para poder organizar los datos.

y obtener el siguiente grafico:



Conclusión A partir de este diagrama de barras podemos ver que el `II CIRCUITO JUDICIAL DE SAN JOSE (GOICOECHEA)`, encabeza y se corona como el circuito mas eficiente a la hora de cerrar casos por violencia domestica.

Pero cual es nuestro top 3 de los mejores así como nuestro top 3 de los peores ?

Circuitos más eficientes:

NombreCircuito	Eficiencia_Carga_Total	Tasa_Resolucion_Pendientes	Reduccion_Pendientes
II CIRCUITO JUDICIAL DE SAN JOSE (GOICOECHEA)	21.17	25.05	6.727950
I CIRCUITO JUDICIAL DE GUA-NACASTE (BAGACES)	19.06	21.32	9.450549
CIRCUITO JUDICIAL DE PUNTARENAS (QUEPOS)	17.99	20.36	7.194245

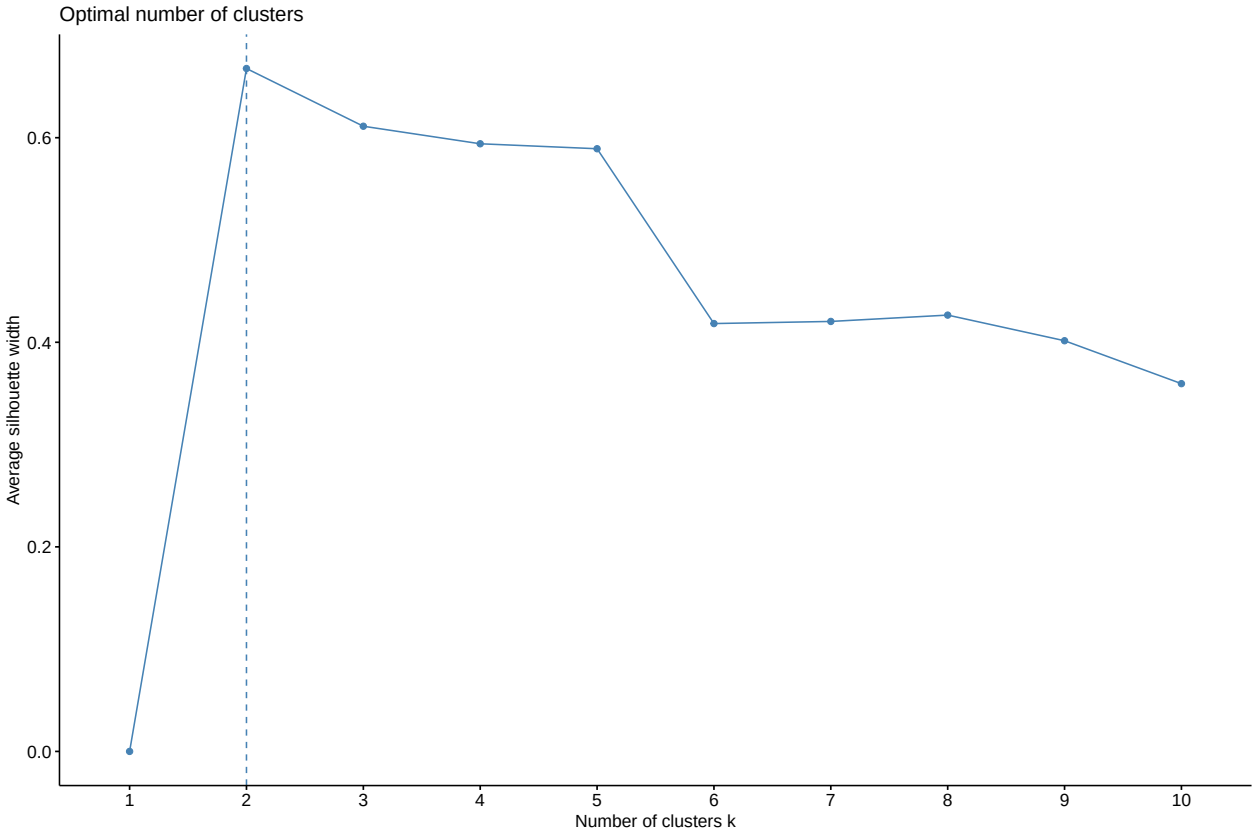
Circuitos menos eficientes:

NombreCircuito	Eficiencia_Carga_Total	Tasa_Resolucion_Pendientes	Reduccion_Pendientes
I CIRCUITO DE LIMÓN (BATÁN-MATINA)	4.11	4.46	-4.199475
II CIRCUITO DE LIMÓN (SIQUIRRES)	4.80	5.29	-4.869857
III CIRCUITO JUDICIAL DE SAN JOSÉ (ASERRÍ)	6.57	6.94	1.339579

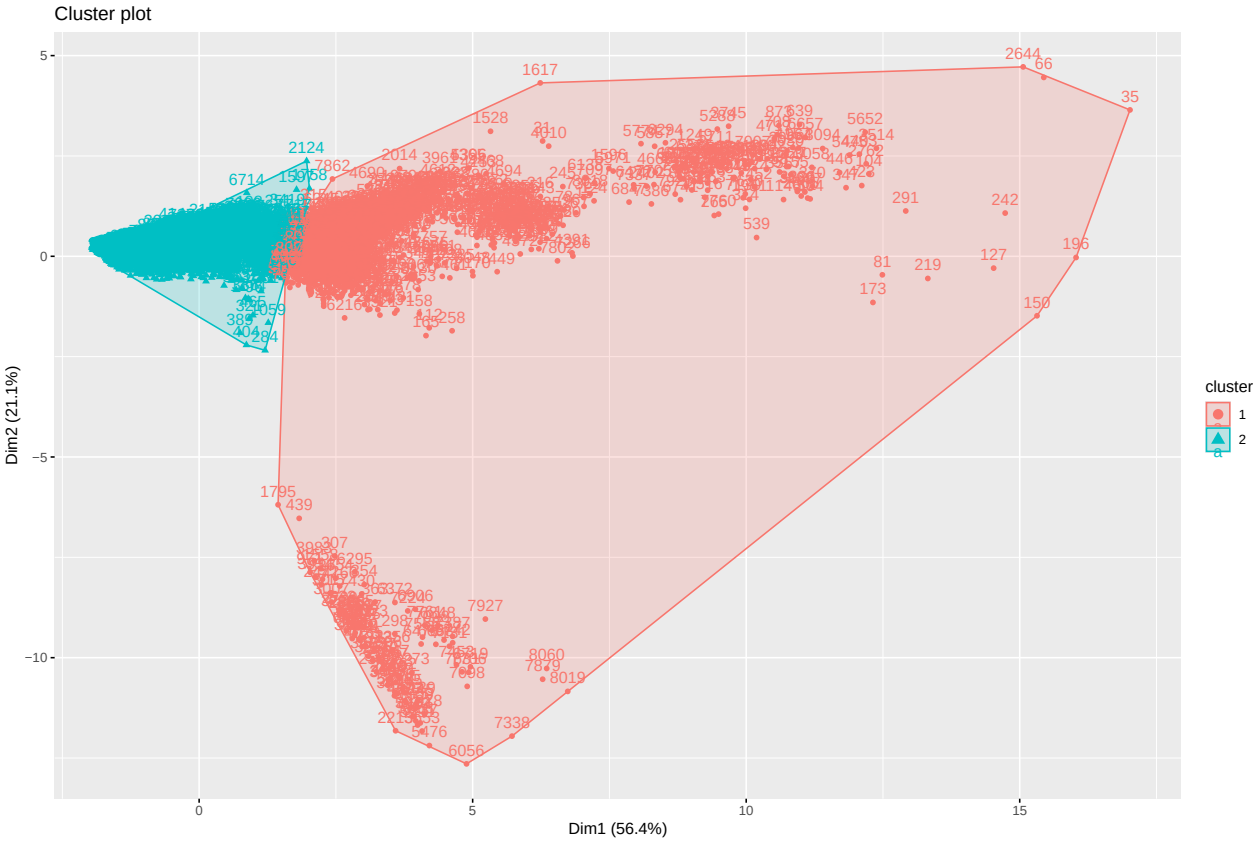
Clustering

Veamos los datos bajo la lupa de un modelo no supervisado (clustering) para entender un poco mas los datos, primero lo queremos ver como un todo utilizando solo las variables nûmericas (no constantes).

Cuantos clusters deberiamos tener?



Según el anterior diagrama deberian ser 2, así que una vez ejecutado nuestro codigo este es el resultado:

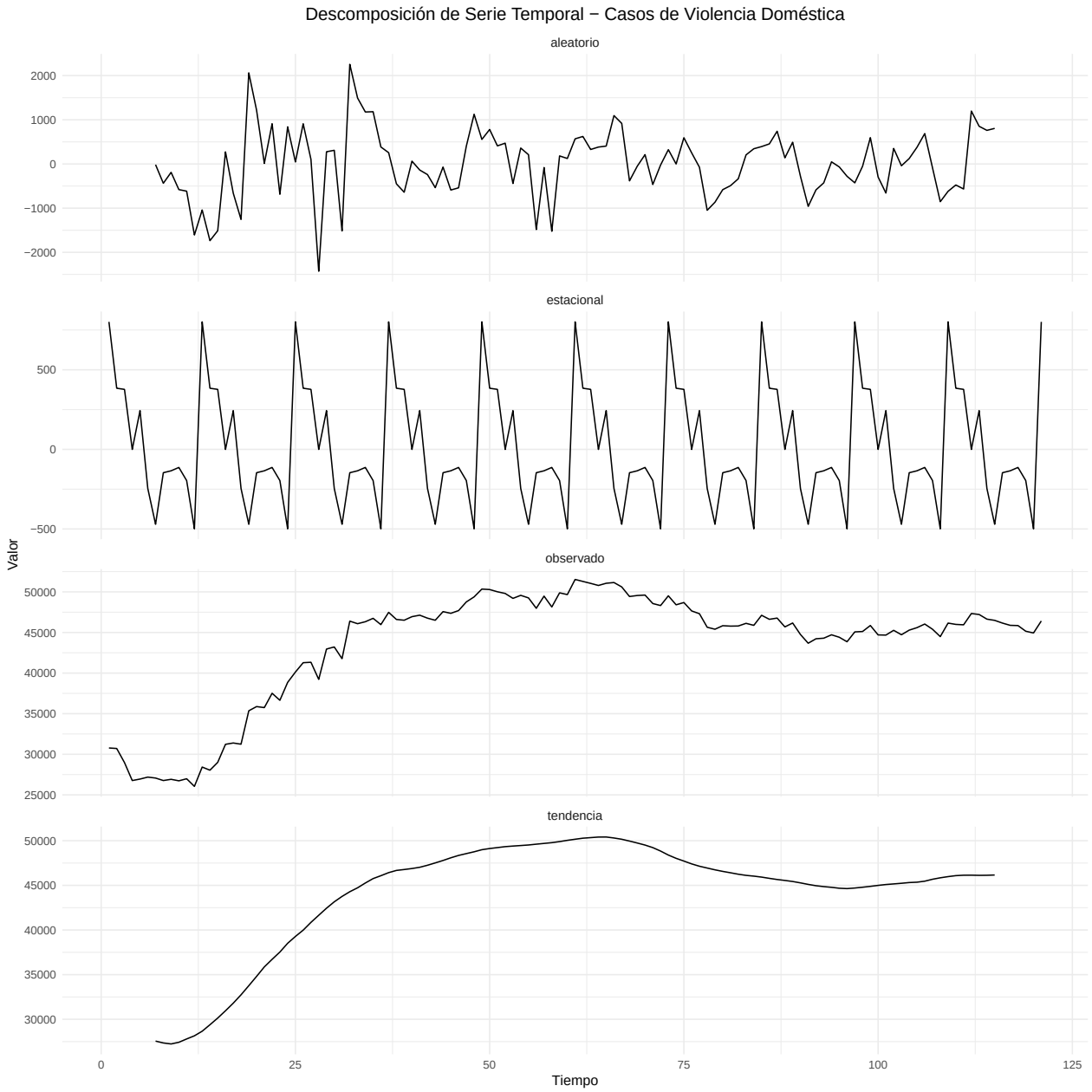


Del siguiente gráfico podemos interpretar lo siguiente:

- Los datos se agrupan naturalmente en dos categorías distintas
- El cluster 1 (rojo) muestra mayor variabilidad y dispersión en ambas dimensiones
- El cluster 2 (turquesa) es más homogéneo y compacto
- La dimensión 1 (eje X) explica aproximadamente el 56.4% de la varianza en los datos
- La dimensión 2 (eje Y) explica aproximadamente el 21.1% de la varianza

Series Temporales (ARIMA)

Dado a que nuestro Conjunto de datos cuenta con meses y años quisimos utilizar un modelo de forcasting de Series temporales llamado ARIMA (Autoregressive Integrated Moving Average) o Medida Movil Integrada Autorregresiva para intentar predecir como se comportará la violencia domestica en nuestro país.



Una breve explicación de lo que se observa:

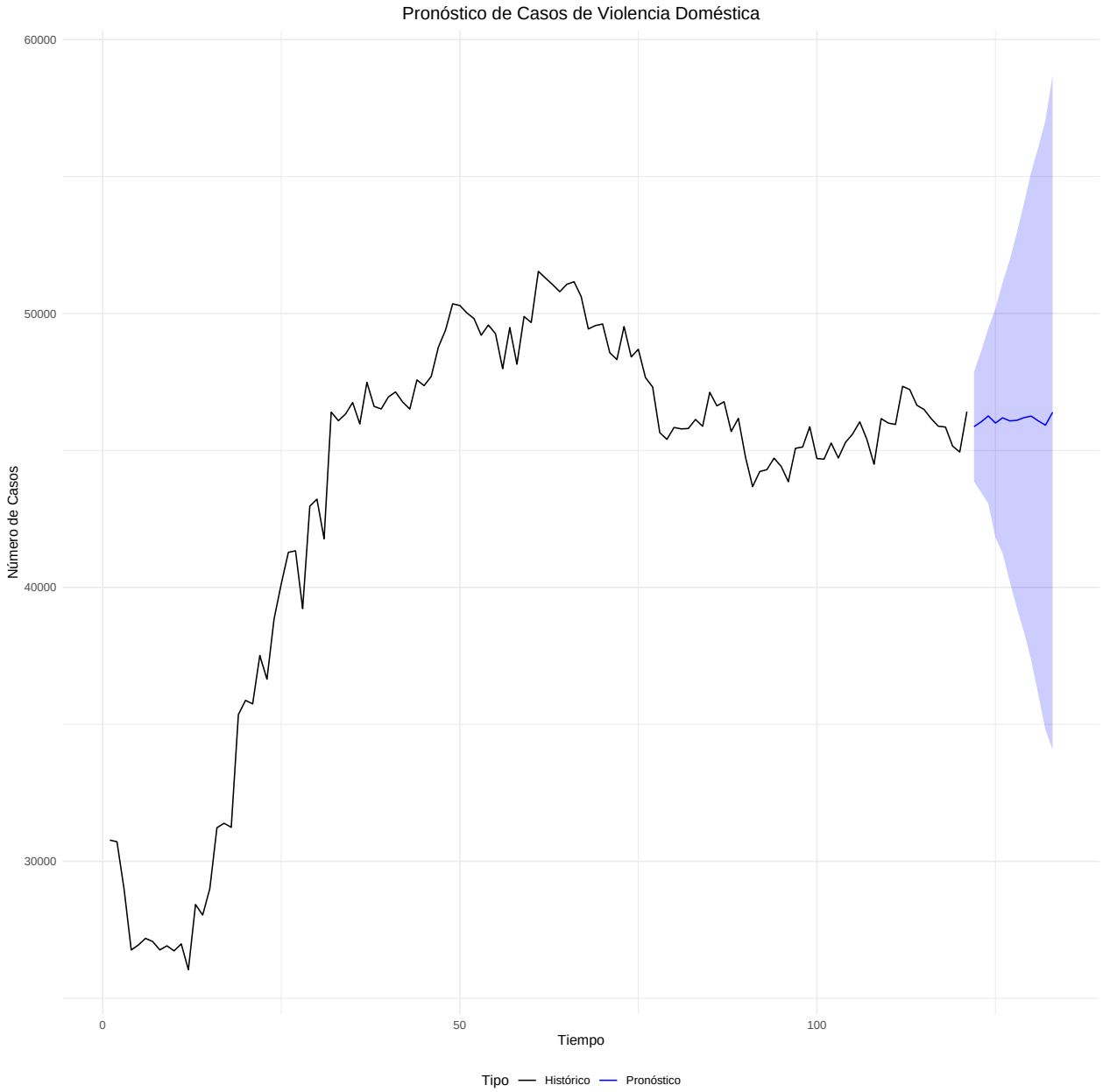
Aleatorio (primer gráfico): Representa la variación irregular o residual que queda después de extraer los componentes de tendencia y estacionalidad. Estos son eventos impredecibles que afectan temporalmente los datos pero no forman parte de ningún patrón sistemático. En este caso, se observan fluctuaciones considerables que oscilan entre aproximadamente +2,000 y -2,000.

Estacional (segundo gráfico): Muestra patrones cíclicos que se repiten a intervalos regulares. El patrón estacional es muy marcado y consistente, con picos pronunciados que se repiten aproximadamente cada 10-12 períodos de tiempo, sugiriendo una variación cíclica (posiblemente mensual o trimestral) en los reportes de violencia doméstica.

Observado (tercer gráfico): Muestra los datos originales sin procesar. Se aprecia un incremento significativo en los casos durante los primeros 50 períodos de tiempo, alcanzando un máximo de aproximadamente 50,000 casos alrededor del período 75, seguido de un ligero descenso y estabilización alrededor de 45,000 casos.

Tendencia (cuarto gráfico): Representa el movimiento a largo plazo de la serie. Muestra un claro aumento desde aproximadamente 28,000 hasta 50,000 casos durante los primeros 75 períodos, seguido de una ligera disminución y posterior estabilización con una leve tendencia al alza hacia el final.

Que podemos esperar en nuestro país ?



A notar del anterior gráfico es que a medida que avanza el tiempo, crece la incertidumbre, el intervalo de confianza se ensancha considerablemente.

Conclusión

Nuestro análisis nos mostró que la violencia doméstica en Costa Rica es un problema complejo que tiene patrones claros según el lugar y el tiempo. Encontramos que las zonas del interior del país, especialmente San José, tienen la mayor cantidad de casos. También vimos que hay más denuncias en los primeros meses del año, lo que sugiere que se necesita más ayuda durante ese tiempo. Al estudiar la eficiencia, descubrimos que algunos circuitos judiciales manejan mejor los casos que otros, lo que nos indica que hay oportunidades para mejorar. Nuestras predicciones sugieren que el problema seguirá siendo un desafío importante en el futuro.

Lo que encontramos puede ser útil para:

- Las personas que toman decisiones en el gobierno
- Los que trabajan en los tribunales

- Los que dan servicios de ayuda social

Esto les puede ayudar a:

- Distribuir mejor los recursos
- Crear mejores estrategias de prevención
- Manejar mejor los casos
- Prepararse para las necesidades futuras

Este proyecto nos permitió aplicar lo aprendido en el curso y ver cómo las herramientas de minería de datos pueden ayudar a entender mejor problemas sociales importantes como la violencia doméstica.