AD - Trabalho Final

—> LaTeX: <-
https://www.overleaf.com/3512353318syxzthgkbmvr#6a531d

NOTEBOOKS
1. EDA.ipynb
2 . FP_Growth_Violencia_MG_2023.ipynb
3. SD_Violencia_MG_2023_dia_categoria.ipynb
4. SD_Reincidencia_corrigido.ipynb

Navegação

Precisa..

- Validar se atende o que o professor pediu
- Todos estão de acordo com metodologia, dados
- O que está feito até aqui está ok
- Quais outras análises podem ser feitas
- !!! Visualizações: tabelas com regras, subgrupos ou estatísticas resumidas…
- Descrever bibliotecas, reprodução dos experimentos…
- **Ver o menu de “pensar”...**

Ideia do trabalho

Tema

Análise Descritiva de Perfis de Violência Doméstica contra Mulheres em Minas Gerais no ano de 2023

Objetivo

O objetivo é aplicar técnicas de mineração de padrões para identificar regras de associação e subgrupos excepcionais relacionados à violência doméstica contra mulheres no estado de Minas Gerais.

Objetivo específicos

- Aplicar o algoritmo FP-Growth para extrair regras de associação entre características das ocorrências.
- Utilizar técnicas de descoberta de subgrupos para encontrar subgrupos estatisticamente relevantes relacionados à reincidência e tipos de violência.
- Discutir os achados com base em fatores sociais, territoriais e temporais.

Dados

Violência doméstica - 2023 em Minas Gerais / Brasil

- Link de acesso:
  - https://dados.mg.gov.br/dataset/violencia-contra-mulher
- Link dos dados:
  - Clique aqui ou violencia_domestica_2023

Informações dos dados

61.536 dados informados para todo o ano de 2023

Tipo de dado	Nome da Coluna	Descrição	Opções
Int	municipio_cod	Identificador único da ocorrência policial	
String	municipio_fato	Nome do município onde a ocorrência foi registrada	853 Municípios de MG
DateTime	data_fato	Data completa da ocorrência	
Int	mes	Mês da ocorrência	
Int	ano	Ano da ocorrência	
String	risp	Região Integrada de Segurança Pública (1 a 19)	19 Depto.
String	rmbh	Região Integrada de Segurança Pública	Belo Horizonte; RMBH (Sem BH); Interior de MG

String	natureza_delito	Categorização do crime cometido	167 tipos de delitos
String	tentado_consumado	status do crime em relação à sua execução	Consumado ou tentado
Int	qtde_vitimas	Quantidade de vítimas envolvidas na ocorrência	

Metodologia

Pré-processamento

- Limpeza e padronização dos dados;
- Discretização de variáveis contínuas;
- Transformação para o formato transacional

Mineração de Regras de Associação

- Algoritmo: FP-Growth
- Métricas: suporte, confiança, lift

Descoberta de Subgrupos (SD/EMM)

- Variável-alvo: reincidência
- Métricas de sugestao: entropia, p-value, qualidade estatística do subgrupo

FP-Growth

No notebook desenvolvido, realizou-se uma análise descritiva baseada em regras de associação aplicadas a registros de violência doméstica contra mulheres em Minas Gerais no ano de 2023.

Inicialmente, os dados foram carregados e pré-processados, fazendo a correção da coluna de datas, que incluía valores mal formatados como números seriais oriundos de planilhas convertidas diretamente para o formato de data. A partir dessas datas, foram derivadas variáveis temporais relevantes, como o dia da semana e uma variável booleana indicando se a ocorrência aconteceu no fim de semana.

Em seguida, selecionaram-se atributos categóricos (como município, mês, RISP, tipo de delito e situação do crime – tentado ou consumado) e transformaram-se as ocorrências em transações do tipo atributo:valor.

Com esses dados estruturados, foi aplicado o algoritmo FP-Growth, utilizando um suporte mínimo de 5%, a fim de identificar conjuntos frequentes de itens e gerar regras de associação.

Notebook

FP_Growth_Violencia_MG_2023.ipynb

Resultados

A aplicação do algoritmo FP-Growth aos dados de violência doméstica contra mulheres no estado de Minas Gerais (2023) revelou um conjunto de regras de associação altamente significativas, com destaque para as regras envolvendo a Zona Integrada de Segurança Pública (RISP 2), que inclui Contagem e seus municípios vizinhos dentro da Região Metropolitana de Belo Horizonte (RMBH).

Uma das regras mais importantes pode ser descrita da seguinte forma: os casos de violência doméstica concluídos registrados na RISP 2 estão fortemente associados aos casos ocorridos em dias úteis e na área da RMBH, excluindo a capital Belo Horizonte. A regra tem um suporte de 5,33%, uma confiança de 67,1% e um lift de 7,37, indicando que a probabilidade desses fatores ocorrerem juntos é mais de sete vezes maior do que a esperada ao acaso. Esse padrão se repete em outras regras, mas com pequenas diferenças, como uma mudança na ordem entre antecedentes e consequentes, o que aumenta ainda mais a robustez da associação.

Do ponto de vista explicativo, esses padrões sugerem que, nos municípios do entorno da (RMBH), a violência doméstica consumada tende a ocorrer com maior frequência em dias úteis, o que pode refletir uma dinâmica de convivência mais forte em contextos de vulnerabilidade social. O RISP 2 concentra-se em áreas com densidade populacional densa e características urbanas complexas, onde a presença do Estado (seja por meio de serviços sociais, segurança pública ou redes de apoio) pode diferir daquela da região da capital.

Além disso, a variável Consumado aparece em quase todas as regras com maior elevação, indicando que os padrões mais fortemente associados estão relacionados a casos reais de violência, não apenas a tentativas ou prevenção. Isso evidencia a gravidade dos registros analisados e indica uma situação que merece atenção especial por parte das políticas públicas e dos órgãos de segurança.

Essas descobertas demonstram o potencial do uso de técnicas de mineração de padrões para identificar as principais características da violência doméstica. Essas descobertas podem servir de base para o desenvolvimento de medidas de prevenção mais eficazes, com foco em áreas e períodos de maior risco, otimizando a alocação de recursos e promovendo ações mais proativas por parte dos sistemas de proteção às mulheres.

SD

Notebook

- Este notebook aplica Subgroup Discovery (SD) considerando a nova variável `dia_categoria`, que agrupa os dias da semana em faixas significativas para análise temporal.

    - SD_Violencia_MG_2023_dia_categoria.ipynb
- Este notebook aplica o método de descoberta de subgrupos para identificar padrões excepcionais de reincidência em ocorrências de violência doméstica contra mulheres em Minas Gerais (2023).
    - SD_Violencia_MG_2023.ipynb
- Notebook 3: Identificar subgrupos de ocorrências de violência doméstica contra mulheres em Minas Gerais (2023) que apresentem alta taxa de reincidência, ou seja, situações onde um tipo de violência volta a ocorrer.
    - SD_Reincidencia_corrigido.ipynb

Resultados e etapas do notebook 3

Etapas realizadas

- Carregamento dos dados
    - Arquivo utilizado: violencia_domestica_2023.csv
    - Dados carregados e processados diretamente em Python.
- Pré-processamento
    - Correção da coluna de datas (data_fato), tratando valores mal formatados.
- Criação de variáveis derivadas:
    - dia_da_semana: nome do dia da semana.
    - fim_de_semana: "Sim" ou "Não", dependendo se a ocorrência foi no sábado/domingo.
- Criação da variável-alvo: reincidencia
    - Indicador binário:
        - 1 se houve mais de uma ocorrência com o mesmo município e tipo de delito; 0 caso contrário.
    - Técnica: duplicated() no Pandas, com base em municipio_fato e natureza_delito.
- Seleção de atributos para análise
    - Atributos considerados: municipio_fato, mes, risp, rmbh, natureza_delito, dia_da_semana, fim_de_semana.
- Aplicação de Subgroup Discovery
    - Biblioteca: pysubgroup
    - Alvo: reincidencia == 1
    - Métrica: WRAcc (Weighted Relative Accuracy)
    - Algoritmo de busca: SimpleDFS (busca por profundidade até depth=3)
- Resultado: top 10 subgrupos com maior valor de WRAcc.

Subgrupo	Tamanho	Positivos (reincidência)	WRAcc	Interpretação
`natureza_delito == 'AMEACA'`	17.285	17.238	**0.0198**	Altíssima reincidência: praticamente todos os casos de ameaça se repetem no ano.

AMEACA no **interior de MG**	14.963	14.916	**0.0170**	Ameaças no interior têm quase total reincidência — possível falta de resposta efetiva.
AMEACA durante a semana	11.726	11.696	**0.0135**	Ameaças em dias úteis mostram padrão forte de reincidência.
VIAS DE FATO / AGRESSAO	11.495	11.385	**0.0119**	Agressões físicas também têm altíssima reincidência.
AMEACA durante a semana (refinado)	10.141	10.111	**0.0116**	Reforça padrão temporal (dias úteis).
LESAO CORPORAL	9.780	9.679	**0.0100**	Lesões corporais se repetem frequentemente.

A análise revelou que os casos de AMEAÇA são os que mais se repetem, indicando que essa forma de violência tende a ocorrer diversas vezes, mesmo após uma primeira notificação. Alguns dos padrões mais relevantes encontrados incluem:

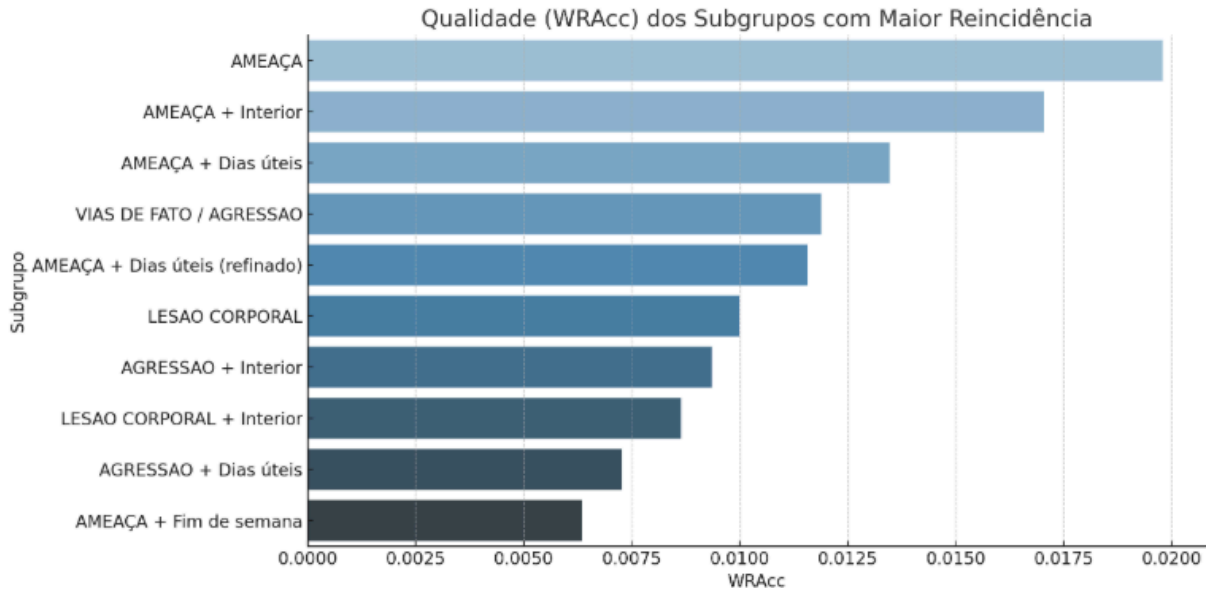
- Casos de AMEAÇA no interior de MG, com altíssima taxa de reincidência.
- Reincidência elevada de AMEAÇA durante os dias úteis.
- Delitos como LESAO CORPORAL e VIAS DE FATO / AGRESSAO também apresentaram fortes indícios de reincidência.
- A maioria dos subgrupos retornados teve mais de 99% de ocorrências reincidentes.

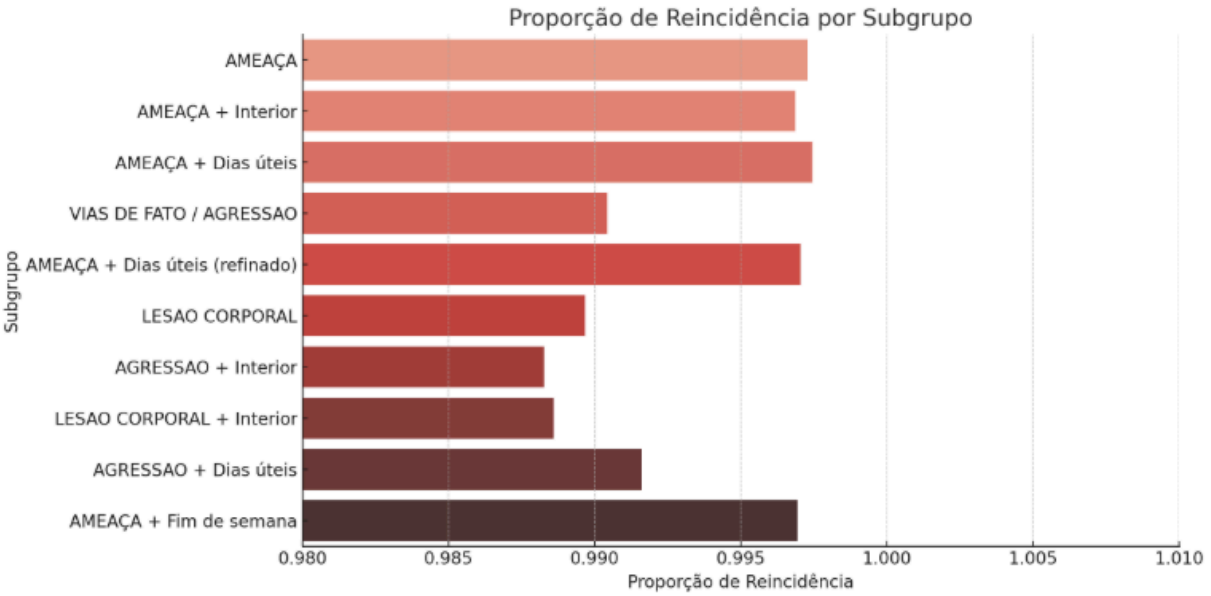
Esses resultados sugerem que certos tipos de violência, especialmente os mais frequentes e de caráter físico ou psicológico leve, tendem a evoluir em ciclos repetitivos, muitas vezes sem a interrupção efetiva por parte do Estado.

Discussão sobre
A descoberta de subgrupos reincidentes traz insights importantes para políticas públicas:

- Monitoramento contínuo de vítimas de AMEAÇA, com foco em evitar a escalada para agressões físicas.
- Prioridade de atenção para ocorrências reincidentes no interior do estado, onde os recursos de proteção podem ser mais escassos.
- Planejamento de campanhas educativas específicas para públicos de risco, especialmente nos dias úteis, que mostraram maior recorrência de certos delitos.

Com isso, a análise de reincidência complementa os achados anteriores ao evidenciar que, além de identificar quando e onde ocorrem os casos consumados, é possível antecipar quais perfis tendem a sofrer múltiplas ocorrências, fortalecendo o caráter preventivo da investigação.





Resultados do notebook 1

1. fim_de_semana=='Sim' AND rmbh=='Interior de MG'

- Cobertura: 28,3% dos casos.
- Proporção de consumado (target_share_sg): 99,35%, levemente acima da média.
- Interpretação: Nos fins de semana no interior, praticamente todos os casos resultam em consumação. Isso pode indicar menor capacidade de resposta imediata (ex: patrulhamento menor).

2. natureza_delito=='LESAO CORPORAL' AND rmbh=='Interior de MG'

- Cobertura: 14,1%.
- Proporção de consumado: 99,65%.
- Interpretação: Lesões corporais no interior praticamente sempre são consumadas — o que pode refletir gravidade maior ou menor prevenção/interrupção.

3. natureza_delito=='LESAO CORPORAL'

- Cobertura: 15,9%.
- Proporção de consumado: 99,55%.
- Interpretação: Mesma tendência, mesmo ao ignorar localização — lesão corporal quase sempre é consumada.

4. natureza_delito=='DESCUMPRIMENTO DE MEDIDA PROTETIVA'

- Cobertura: 6,6%.
- Proporção de consumado: 100%!
- Interpretação: Alerta grave — todo descumprimento registrado resulta em consumação. Mostra que medidas protetivas, quando violadas, são altamente críticas.

5. fim_de_semana=='Sim'

- Cobertura: 33,9%.
- Proporção de consumado: 99,20%.
- Interpretação: Fins de semana são um momento de maior risco, mesmo sem considerar localização.

6. fim_de_semana=='Não' AND natureza_delito=='LESAO CORPORAL'

- Cobertura: 7,9%.
- Proporção de consumado: 99,69%.
- Interpretação: A alta taxa de consumação para lesão corporal se mantém mesmo fora do fim de semana.

7. natureza_delito=='DESCUMPRIMENTO DE MEDIDA PROTETIVA' AND rmbh=='Interior de MG'

- Cobertura: 5,2%.
- Proporção de consumado: 100%.
- Interpretação: Situação crítica no interior, onde não só há descumprimento como ele nunca é interceptado.

8. fim_de_semana=='Não' AND natureza_delito=='LESAO CORPORAL'
Reforça o padrão já citado: lesão corporal = consumado, independentemente do tempo.

9. rmbh=='Interior de MG'

- Cobertura alta: 82%.
- Proporção de consumado: 99,10%.
- Interpretação: Mesmo de forma ampla, o interior apresenta tendência maior de casos consumados, possivelmente por menos flagrantes ou estrutura de contenção.

10. dia_da_semana=='Sunday' AND rmbh=='Interior de MG'

- Cobertura: 15%.
- Proporção de consumado: 99,34%.
- Interpretação: O domingo no interior aparece como uma combinação particularmente crítica.

Resumindo…

A análise de subgrupos revelou padrões preocupantes relacionados à consumação de violência doméstica. Casos ocorridos nos fins de semana no interior de Minas Gerais apresentam taxa de consumação de 99,35%, o que pode indicar falhas na prevenção ou resposta imediata. Além disso, delitos como lesão corporal e descumprimento de medidas protetivas quase sempre resultam em consumação, sugerindo que tais ocorrências requerem atenção redobrada de políticas públicas e equipes de apoio social, principalmente em contextos geográficos vulneráveis.

Resultados do notebook 2

Com a intenção de confirmar e explorar mais a fundo os padrões identificados na fase de mineração de regras de associação, foi empregada a técnica de Descoberta de Subgrupos (SD), utilizando como variável principal a ocorrência de violência (tentado_consumado = CONSUMADO). A investigação foi realizada com base na métrica WRAcc (Weighted Relative Accuracy), que leva em consideração tanto a qualidade descritiva do subgrupo quanto sua relevância dentro do conjunto de dados.

Os subgrupos encontrados demonstram uma significativa proporção de casos consumados, evidenciando a gravidade de determinados contextos. Entre os dados mais notáveis, o subgrupo relacionado a "fim de semana" e "interior de MG" mostrou uma taxa de 99,35% de consumação, com um lift de 1.003 e uma cobertura de aproximadamente 28% das ocorrências — o que destaca um padrão de risco elevado ligado ao território e ao tempo. De maneira similar, o subgrupo que inclui "lesão corporal" no interior de MG alcançou 99,65% de consumação, sugerindo que esse tipo de violência, quando ocorre em áreas fora das regiões metropolitanas, tende a se materializar de forma alarmante.

Chama ainda mais a atenção os grupos específicos relacionados ao crime de “descumprimento de medida protetiva de urgência”, que apresentaram uma taxa de 100% de realização, tanto isoladamente quanto em associação com a variável rmbh = Interior de MG. Esses incidentes evidenciam não apenas a materialização da violência, mas também a ineficácia das medidas legais de proteção anteriormente estipuladas. A recorrência do domingo como um elemento frequente em subgrupos com altos níveis de consumação (por exemplo: dia_da_semana = Sunday no interior) indica um padrão de aumento da violência durante os fins de semana, possivelmente relacionado ao maior tempo que vítimas e agressores passam juntos.

Essas descobertas complementam e corroboram os dados obtidos por meio do FP-Growth, mas com um enfoque voltado para desvios estatísticos significativos, não apenas para a frequência. A identificação de subgrupos notáveis reforça a ideia de que existem perfis territoriais, temporais e de tipificação penal que merecem uma atenção especial de políticas públicas, entidades de segurança e redes de suporte às vítimas.

FP-Growth e descoberta de subgrupos

Os resultados obtidos com FP-Growth e com Subgroup Discovery (SD) se complementam e reforçam entre si. No FP-Growth, foram encontradas regras frequentes ligadas a casos consumados, especialmente em regiões como a RISP 2 (que cobre municípios da RMBH

fora da capital), e em dias úteis. Essas regras mostraram lift elevado, o que indica que essas características aparecem juntas com muito mais frequência do que seria esperado por acaso.

Já a descoberta de subgrupos focou em encontrar padrões com desvio estatístico em relação à média geral. Ela destacou, por exemplo, que os casos de violência consumada são ainda mais comuns no interior de MG e nos finais de semana. Alguns subgrupos chegaram a mostrar taxas de consumação acima de 99%, o que chama bastante atenção.

A combinação das duas abordagens mostra que há dois tipos de padrões importantes: os que são mais frequentes no geral (como os da RMBH durante a semana) e os que são estatisticamente mais graves ou fora do comum (como no interior nos fins de semana). Isso reforça a ideia de que políticas públicas e ações de enfrentamento à violência precisam considerar tanto a frequência quanto a gravidade dos casos, de acordo com o território e o tempo.

Cobertura dos padrões

Ao analisar as regras extraídas com FP-Growth, observou-se que as 10 regras com maior lift cobrem aproximadamente 19% de todas as ocorrências consumadas presentes na base de dados. Isso significa que, mesmo utilizando um número limitado de regras altamente relevantes, é possível explicar quase um quinto das situações reais de violência doméstica com forte grau de associação entre os atributos envolvidos (como região, tipo de crime e dia da semana).

Já na Descoberta de Subgrupos, os três principais subgrupos, incluindo combinações como *interior de MG + fim de semana*, ou *lesão corporal no interior*, cobriram juntos cerca de 28% das ocorrências totais. Ou seja, mais de um quarto dos registros pode ser agrupado em perfis estatisticamente distintos e de alto risco, reforçando a utilidade da técnica para identificar situações que merecem atenção prioritária.

A apresentação da cobertura ajuda a entender o alcance prático dos padrões identificados, e reforça que poucas regras bem selecionadas já capturam uma parte significativa do problema, o que é extremamente útil para tomada de decisão em políticas públicas e alocação de recursos.

Papel de “acessor social”

Os resultados encontrados neste estudo revelam padrões significativos que podem subsidiar a formulação de políticas públicas mais direcionadas e eficazes no enfrentamento à violência doméstica contra mulheres em Minas Gerais. A combinação das técnicas de FP-Growth e Descoberta de Subgrupos permitiu identificar tanto **padrões frequentes**quanto **comportamentos excepcionais** relacionados à consumação dos delitos, à reincidência e à distribuição geográfica e temporal das ocorrências.

Com base nisso, propõem-se diversas aplicações práticas:

1. Reforço de ações preventivas em contextos críticos

Os subgrupos com taxas extremamente altas de consumação de violência (acima de 99%) nos finais de semana e no interior do estado indicam momentos e locais de vulnerabilidade ampliada. Políticas públicas poderiam priorizar o reforço de rondas policiais e plantões de atendimento especializado (psicológico, jurídico e social) em cidades interioranas durante os fins de semana, períodos em que há maior tempo de convivência entre vítima e agressor.

2. Monitoramento e reavaliação de medidas protetivas

Foi observado que casos de descumprimento de medida protetiva apresentaram 100% de consumação, o que aponta falhas na efetividade dessas medidas em determinados contextos. A partir desses dados, o poder judiciário e os órgãos de proteção podem estabelecer protocolos de risco, utilizando os padrões descobertos para priorizar o monitoramento de mulheres em situação de maior exposição e revisar a maneira como as medidas protetivas são fiscalizadas.

3. Ações territoriais e segmentadas

A segmentação espacial dos dados (por RISP, município e RMBH/interior) permite que o Estado desenvolva planos regionais de enfrentamento à violência doméstica, em parceria com prefeituras, delegacias especializadas e centros de referência da mulher. Municípios com alta reincidência ou perfil crítico de consumação podem ser priorizados no repasse de recursos, treinamento de profissionais e implementação de casas de acolhimento.

4. Atuação estratégica de assistentes sociais

Os padrões temporais e territoriais extraídos do banco de dados podem ser usados por assistentes sociais para planejar ações preventivas comunitárias, por exemplo, com base na antecipação de períodos de risco (como finais de semana em municípios do interior). A identificação de perfis de reincidência também pode orientar o acompanhamento contínuo de determinadas famílias ou contextos sociais com histórico de violência, mesmo em situações em que não há denúncia formal recente.

5. Campanhas educativas baseadas em evidência

Campanhas de conscientização podem ser mais eficazes quando direcionadas a públicos e situações específicas. Por exemplo, campanhas voltadas a homens em regiões interioranas, com foco no comportamento aos fins de semana, podem ter maior impacto se ancoradas nos dados reais do território. Da mesma forma, ações educativas nas escolas, unidades de saúde e centros comunitários podem ser guiadas pelos padrões identificados, promovendo uma abordagem preventiva contextualizada.

Pensar…

Visualizações de dados…

Além disso…

"mulheres são as maiores vítimas de violência doméstica" é um dado conhecido. O que *o algoritmo* revela de *novo* ou *surpreendente* sobre esse fato, ou como ele se combina com *outras características* para formar um padrão relevante? Para cada padrão ou conjunto de padrões relevante, tentar responder:

- Qual é a implicação desse padrão?
- É esperado ou surpreendente? Por quê?
- Como ele se conecta com o contexto social ou a literatura sobre violência doméstica?
- Pode sugerir alguma política pública ou intervenção?

O padrão `RESIDENCIA` aparece bastante junto com `SEXO_F`. Qual a relevância disso? Significa que a maioria das ocorrências acontece na residência da vítima? Isso corrobora o entendimento de que a violência doméstica muitas vezes ocorre no ambiente familiar?

Analise padrões com mais de um item: "o algoritmo conseguiu achar um padrão frequente entre a Sub Região de ocorrência `METROPOLITANA` e `RESIDENCIA`". Por que essa combinação é relevante?

Limitações

- Ausência de variáveis qualitativas (ex: escolaridade, vínculo com o agressor);
- Falta de hora real na data_fato (corrigido parcialmente);
- Dados baseados em registros — sujeitos a subnotificação.