UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE ARTES E COMUNICAÇÕES

DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO

CURSO DE GRADUAÇÃO EM GESTÃO DA INFORMAÇÃO

ANDRÉ FELIPE CICCO RIBAS

DESCOBERTA E ANÁLISE DE PADRÕES NA BASE DE DADOS DE INFRAÇÕES DE TRÂNSITO DO RECIFE

Recife

2025

ANDRÉ FELIPE CICCO RIBAS

DESCOBERTA E ANÁLISE DE PADRÕES NA BASE DE DADOS DE INFRAÇÕES DE TRÂNSITO DO RECIFE

Trabalho de Conclusão de Curso apresentado ao Curso de Gestão Informação da Universidade Federal de Pernambuco – UFPE, como requisito parcial para obtenção do grau de Bacharel em Gestão da Informação.

Orientador (a): Bruno Tenório Ávila

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Ribas, André Felipe Cicco.
Descoberta e análise de padrões na base de dados de infrações de trânsito do Recife / André Felipe Cicco Ribas. - Recife, 2025.
70 p. : il., tab.

Orientador(a): Bruno Tenório Ávila
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Artes e Comunicação, Gestão da Informação - Bacharelado, 2025.
Inclui referências, apêndices.

1. Padrões. 2. Mineração de dados. 3. Enriquecimento de dados. I. Ávila, Bruno Tenório. (Orientação). II. Título.

000 CDD (22.ed.)

Serviço Público Federal
Universidade Federal de Pernambuco
Centro de Artes e Comunicação
Departamento de Ciência da Informação

FOLHA DE APROVAÇÃO

DESCOBERTA E ANÁLISE DE PADRÕES NA BASE DE DADOS DE INFRAÇÕES DE TRÂNSITO DO RECIFE

ANDRÉ FELIPE CICCO RIBAS

Trabalho de Conclusão de Curso submetido à Banca Examinadora, apresentado no Curso de Gestão da Informação, do Departamento de Ciência da Informação, da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Gestão da Informação.

TCC aprovado em 04 de abril de 2025

Banca Examinadora:

_____

BRUNO TENÓRIO ÁVILA - Orientador(a)
Universidade Federal de Pernambuco - DCI

_____

MÁRCIO HENRIQUE WANDERLEY FERREIRA – Examinador(a) 1
Universidade Federal de Pernambuco - DCI

_____

MARCÍLIO BEZERRA CRUZ - Examinador(a) 2
Doutorando –PPGCI/UFPE

DEDICATÓRIA

Dedico este trabalho à minha família, alicerce da minha vida, por todo o amor, apoio e incentivo ao longo da minha jornada. Aos meus pais, por cada gesto de amor, por cada palavra de incentivo. Por me ensinarem, através do exemplo, que os objetivos são alcançados com esforço, paciência e dignidade. Este trabalho é fruto do suporte que recebi, pois sem o suporte emocional, as renúncias e a confiança que depositaram em mim, eu não teria chegado até aqui.

Ao meu avô, Fernando e minhas avós Tereza, que me ensinaram o valor das pequenas coisas e humildade, e por todo carinho. Dedico também a minha irmã, que tenho como referência por todo apoio e incentivo e palavras de conforto durante o processo e na vida.

AGRADECIMENTOS

Agradeço à minha família, minha base e referência. Obrigado por estarem sempre ao meu lado, torcendo, incentivando e, acima de tudo, acreditando em mim. Aos meus pais, pelo amor incondicional, pelos conselhos e por todos os sacrifícios silenciosos que me trouxeram até aqui. Sem o apoio de vocês, nada disso teria sido possível.

Chegar até aqui foi uma trajetória repleta de desafios, aprendizados e superações. Por isso, expresso minha mais sincera gratidão a todos que, de alguma forma, fizeram parte deste processo tão significativo da minha vida. A Deus, por me dar forças nos momentos mais difíceis, por iluminar meu caminho e esperança mesmo quando tudo parecia incerto.

As minhas avós Tereza e avô Fernando, que durante os estudos perguntava se o rádio estava alto, cuja força, humildade e carinho deixam marcas profundas no meu coração. Agradeço também a minha irmã Alessandra por todos os ensinamentos, pelos apontamentos e conselhos. À Ayra, minha fiel companheira de quatro patas, presença constante nas noites de estudo, cuja companhia silenciosa tornou cada momento mais leve. A vocês, minha eterna gratidão. Meu agradecimento vai também para Marília, pelo apoio constante, incentivo e, sobretudo, pelo companheirismo durante todo esse processo.

Aos professores, em especial ao professor Bruno que compartilhou não apenas conhecimento, mas também me orientou, ofereceu oportunidades, tornou possível esse trabalho e que tenho como referência. Obrigado por acreditarem no meu potencial e por contribuírem para a minha formação com tanta generosidade.

Aos colegas e amigos que caminharam ao meu lado durante essa jornada, dividindo dúvidas e conquistas. Ter vocês por perto tornou tudo mais leve e significativo.

A todas as pessoas que, direta ou indiretamente, contribuíram para que este trabalho fosse concluído: meu muito obrigado.

RESUMO

Este trabalho investiga como a mineração de dados públicos sobre infrações de trânsito pode revelar padrões de comportamento na cidade do Recife. Utilizando dados do Portal de Dados Abertos do Recife, a pesquisa adota uma abordagem exploratória e descritiva, com o objetivo de descobrir e analisar os padrões identificados. A metodologia envolve a coleta e a mineração de dados, por meio do uso do algoritmo Apriori, com o intuito de identificar associações relevantes entre os registros de infrações de trânsito e variáveis temporais. Os resultados demonstram que as variáveis temporais influenciam significativamente os padrões de registros de infrações, possibilitando uma melhor compreensão das dinâmicas urbanas e contribuindo para o planejamento estratégico. Conclui-se que a mineração de dados pode colaborar para uma gestão urbana mais eficiente, promovendo transparência, conscientização e fornecendo subsídios para a formulação de políticas públicas.

Palavras-chave: padrões; enriquecimento de dados; mineração de dados.

ABSTRACT

This study investigates how the mining of public data on traffic violations can reveal behavioral patterns in the city of Recife. Using data from the Recife Open Data Portal, the research adopts an exploratory and descriptive approach, aiming to discover and analyze identified patterns. The methodology involves the collection and mining of data through the use of the Apriori algorithm, with the objective of identifying relevant associations between traffic violation records and temporal variables. The results show that temporal variables significantly influence the patterns of recorded violations, enabling a better understanding of urban dynamics and contributing to strategic planning. It is concluded that data mining can support more efficient urban management by promoting transparency, raising public awareness, and providing a solid foundation for the development of public policies.

Keywords: patterns; data mining; data enrichment.

LISTA DE QUADROS / TABELAS

Tabela 1 – Exemplo de estrutura de uma tabela de transações. 18
Tabela 2 - Representação da tabela apriori_itens populada 29
Tabela 3 - Representação da tabela apriori_transacoes populada 30
Tabela 4 - Atributos enriquecidos 31

Quadro 1 - Detalhamento dos padrões descobertos..... 33

LISTA DE GRÁFICOS / FIGURAS

Figura 1 - Processo de descoberta de conhecimento em bases de dados 15
Figura 2 - Suporte e confiança 19
Figura 3 - Classificação de regras de associação com base em suporte e confiança 19
Figura 4 - Modelo lógico das tabelas do algoritmo Apriori 29
Figura 6 - Grafo de associação dos padrões descobertos 35

Gráfico 1 - Distribuição das multas por hora e ano 36
Gráfico 2 – Distribuição normalizada das multas por hora e ano 37
Gráfico 3 - Distribuição das multas por hora e mês 39
Gráfico 4 - Distribuição normalizada das multas por hora e mês 40
Gráfico 5 - Distribuição das multas por hora, dias úteis da semana e ano 41
Gráfico 6 - Distribuição normalizada das multas por hora, dias úteis da semana e ano 42
Gráfico 7 - Distribuição das multas ocorridas no sábado por hora e ano 43
Gráfico 8 - Distribuição normalizada das multas ocorridas no sábado por hora e ano 43
Gráfico 9 - Distribuição das multas ocorridas no domingo por hora e ano 45
Gráfico 10 - Distribuição normalizada das multas ocorridas no domingo por hora e ano 45
Gráfico 11 - Regressão linear simples das multas por mês 47
Gráfico 12 - Distribuição das infrações de velocidade acima de 20% por hora e ano 48
Gráfico 13 - Distribuição normalizada das infrações de velocidade acima de 20% por hora e ano 49
Gráfico 14 - Distribuição de infrações de velocidade acima de 20% por hora e mês 51
Gráfico 15 - Distribuição normalizada de infrações de velocidade acima de 20% por hora e mês 52
Gráfico 16 - Distribuição de infrações de velocidade acima de 20% por hora, dia útil e ano 53
Gráfico 17 - Distribuição normalizada de infrações de velocidade acima de 20% por hora, dia útil e ano 54
Gráfico 18 - Distribuição de infrações de velocidade acima de 20% por hora, final de semana e ano 55
Gráfico 19 - Distribuição de infrações de velocidade acima de 20% por hora, final de semana e ano 56

SUMÁRIO

1 INTRODUÇÃO 8

1.1 Objetivos 11

1.2 Visão geral do documento 11

2 REFERENCIAL TEÓRICO 12

2.1 Padrões 12

2.2 Descoberta de conhecimento em base de dados 14

2.2.1 Seleção 15

2.2.2 Pré-processamento dos dados 15

2.2.3 Mineração de dados 16

2.2.3.1 Regras de associação 17

2.2.3.2 Algoritmo Apriori 20

2.2.4 Pós-processamento 21

2.3 Enriquecimento de dados 21

3 PROCEDIMENTOS METODOLÓGICOS 24

3.1 Coleta de dados 24

3.2 Pré-processamento 25

3.3.1 Transformação dos dados 25

3.3.2 Implementação do algoritmo Apriori 27

3.3 Mineração de dados 31

3.4 Pós-processamento 33

4 RESULTADOS EXPERIMENTAIS 35

4.1 Padrões descobertos 36

5 CONSIDERAÇÕES FINAIS 57

REFERÊNCIAS BIBLIOGRÁFICAS 58

APÊNDICE A – CÓDIGO PYTHON USADO PARA PROCESSAR OS DADOS 63

1 INTRODUÇÃO

No Brasil e no mundo, iniciou-se um processo de rápida convergência tecnológica impulsionado pela ampla disseminação da internet e de suas diversas aplicações. Atualmente, atividades como falar ao telefone, movimentar saldos bancários, verificar multas de trânsito pela internet, trocar mensagens com pessoas de qualquer parte do mundo e realizar pesquisas ou estudos tornaram-se práticas cotidianas em diversos lugares, incluindo o Brasil (Takahashi, 2000). Esse cotidiano só foi possível graças ao avanço acelerado das tecnologias da informação que transformaram profundamente a maneira como as pessoas acessam e utilizam informações.

Com o progresso dessas tecnologias, vive-se em uma era marcada pela produção e circulação de um volume crescente de informações, denominada por Takahashi (2000) como *sociedade da informação*. Na qual, Lastres (1999) aponta que a informação, assume valores sociais e econômicos importantes. Nesse contexto, o uso e o acesso aos dados gerados se tornam essenciais para a inovação e a eficácia nas políticas públicas (Kučera, Chlapek e Nečaský, 2013). Para que essas informações sejam utilizadas efetivamente na melhoria dos serviços públicos, é imprescindível não apenas garantir o acesso a esses dados, mas também promover a qualidade das informações para uma análise adequada.

No Brasil, o fortalecimento da transparência e do acesso às informações públicas se deu por meio de legislações que fomentam a disponibilização e o uso desses dados de forma aberta e acessível. A Lei nº 12.527/2011, conhecida como Lei de Acesso à Informação (LAI), é um dos pilares dessa mudança, porque regulamenta o direito de qualquer cidadão a obter informações dos órgãos e entidades públicas, promovendo a transparência governamental e a participação cidadã (Conradie; Choenni, 2014; Evans; Campos, 2013).

A LAI determina que os dados públicos sejam disponibilizados de forma transparente, acessível e reutilizável, com linguagem clara e formatos abertos. Sua implementação estimulou a criação de diversos portais de dados abertos no Brasil, como o Portal de Dados Abertos do Recife, que oferece informações em áreas como saúde, educação, mobilidade, infraestrutura e segurança, promovendo o controle social e a transparência governamental.

Complementando a LAI, o Decreto nº 8.777, de 11 de maio de 2016, estabeleceu a Política de Dados Abertos do Poder Executivo Federal, com o intuito de aprimorar a cultura de transparência pública e consolidar a obrigatoriedade de que os órgãos da administração pública federal disponibilizassem suas bases de dados de maneira acessível, em formatos reutilizáveis, promovendo a integração de informações e a criação de soluções inovadoras (Brasil, 1997).

Essa política visa não apenas a transparência, mas também a eficiência, permitindo que cidadãos, empresas e instituições independentes possam utilizar os dados públicos para pesquisas. Além de estabelecer diretrizes para o acesso à informação, a regulamentação brasileira também influencia diretamente a forma como os dados de infrações de trânsito são coletados, organizados e disponibilizados. O Código de Trânsito Brasileiro (Lei nº 9.503/1997), por exemplo, determina a obrigatoriedade do registro e sistematização das infrações cometidas, cabendo aos órgãos executivos de trânsito a responsabilidade pela sua gestão e transparência (Brasil, 1997).

Com a criação da Política de Dados Abertos (Decreto nº 8.777/2016), reforça-se o papel das instituições públicas como a Autarquia de Trânsito e Transporte Urbano de Recife (CTTU) em divulgar esses registros de maneira padronizada e acessível, contribuindo para a fiscalização social e para o desenvolvimento de estudos acadêmicos e técnicos que busquem compreender o comportamento infracional e propor soluções preventivas. A disponibilização de dados públicos é um passo essencial para alcançar o conceito de cidades inteligentes. Segundo a Carta Brasileira para Cidades Inteligentes, essas cidades buscam promover o desenvolvimento sustentável, a inclusão social, a inovação e uma governança colaborativa (Brasil, 2022, p. 28).

A partir da análise de dados públicos sobre infrações de trânsito disponibilizados pela CTTU, pode-se identificar possíveis padrões latentes que ajudam a entender o comportamento urbano de forma mais sistemática e fundamentada. Esse tipo de análise aproxima-se da noção de padrões proposta por Alexander (1977).

Segundo Netto e Krafta (2009), compreender o comportamento urbano demanda uma abordagem que ultrapasse os dos métodos quantitativos tradicionais. É necessário adotar métodos, capazes de captar os efeitos que a forma urbana exerce sobre as dinâmicas sociais, econômicas e ambientais das cidades. Esses efeitos podem se manifestar em padrões estruturais, como por exemplo a distribuição de densidade, acessibilidade, centralidades e fluxos, os quais impactam diretamente nas escolhas de mobilidade dos indivíduos. De acordo com Hillier (2007), a configuração do espaço é uma condição fundamental da vida social.

Conforme Batty (2007) e Bettencourt (2010), as cidades funcionam como sistemas adaptativos complexos, compostos por múltiplos agentes interagindo localmente e produzindo padrões emergentes que só podem ser revelados por meio de análises estatísticas e espaciais.

É neste contexto que surge o problema da pesquisa: **como os dados públicos de infrações de trânsito do Recife podem ser analisados para descobertas de padrões?**

Anualmente, um número significativo de infrações de trânsito é registrado por órgãos fiscalizadores em diversas regiões do Brasil, o que evidencia uma instabilidade recorrente no

comportamento dos condutores nas vias urbanas (Jadejiski, 2020). Essas infrações refletem diretamente os modos de comportamento no tráfego, sendo influenciadas por múltiplos fatores, como o tipo de veículo, a infraestrutura viária e a percepção de risco dos motoristas (Feitosa, 2010). De acordo com Rozestraten (1998), erros e violações de trânsito são fenômenos inerentes ao ambiente. Yagil (2001) complementa que erros e violações são fenômenos frequentes. Diante desse cenário, este trabalho contribuirá para usar o processo de descoberta de conhecimento em bases de dados para impulsionar melhorias sociais na cidade, como aprimorar o planejamento estratégico, aprimorar a transparência e conscientização da população relacionados mobilidade urbana na cidade.

Este trabalho também se justifica pela oportunidade de demonstrar como bases de dados podem ser exploradas de maneira mais profunda. Por meio de técnicas de Mineração de Dados e descoberta de padrões, busca-se ampliar o valor informacional ao identificar relações ocultas, tendências e comportamentos nos atributos disponíveis bases públicas oferecidas pelo portal de dados abertos do Recife.

De acordo com o Índice de Dados Abertos para Cidades 2023, que avalia a disponibilidade e a qualidade dos dados abertos nas capitais brasileiras sob uma perspectiva cívica, Recife se posiciona como a terceira capital brasileira no *ranking* geral. No que se refere especificamente à abertura de dados sobre Mobilidade e Transporte Público, a capital pernambucana figura em segundo lugar entre os municípios brasileiros. Além disso, Recife se destaca pela elevada qualidade dos dados relacionados à fiscalização e ocorrências (*Open Knowledge* Brasil, 2024).

Essa abordagem, quando aliada a análise sistemática dessas informações por meio de processos de descoberta de conhecimento em base de dados, tem o potencial de contribuir para uma compreensão mais aprofundada dos dados disponíveis, bem como, contribuir para a descoberta de conhecimento, desafios e oportunidades enfrentados pela cidade. Esse entendimento pode subsidiar a tomada de decisões mais informadas e promover a implementação de políticas públicas mais eficazes.

A crescente disponibilização de dados públicos, impulsionada pela adoção de políticas de governo aberto e transparência a partir da LAI, tem gerado uma grande quantidade de informações valiosas sobre diversos aspectos da sociedade e da administração pública. No entanto, limitações no formato e na qualidade com que as bases de dados são disponibilizadas podem dificultar sua análise e utilização efetiva. Essa dificuldade impede que a sociedade, pesquisadores e até mesmo gestores públicos aproveitem plenamente o potencial informacional desses

dados para a tomada de decisões fundamentadas, o monitoramento de políticas públicas e a geração de novos conhecimentos.

Assim, torna-se essencial a implementação de técnicas de enriquecimento de dados que permitam não apenas consolidar as informações, mas também melhorar sua qualidade e acessibilidade. Este trabalho justifica-se socialmente em entender como essa técnica amplia a análise desses dados públicos, contribuindo para um maior acesso à informação, transparência governamental e, consequentemente, novos conhecimentos para uma cidadania mais ativa e informada.

1.1 Objetivos

O objetivo principal do trabalho é descobrir e analisar os padrões sobre infrações de trânsito do Recife, aplicando técnicas de mineração e enriquecimento de dados em bases de dados disponibilizados pelo Portal de Dados Abertos do Recife.

Os objetivos específicos deste trabalho são:

a) Descrever o processo de coleta e transformação, aplicando métodos de enriquecimento de dados;
b) Implementar técnicas de mineração de dados, utilizando o algoritmo Apriori, explorando os dados e identificando padrões ocultos e associações significativas;
c) Estruturar o conhecimento obtido por meio de padrões.

1.2 Visão geral do documento

Na próxima seção, serão abordados conceitos fundamentais e uma revisão bibliográfica sobre padrões, mineração de dados e enriquecimento de dados, incluindo suas metodologias e ferramentas relevantes para a pesquisa. Essa revisão fornecerá uma base teórica sólida para entender como essas técnicas podem ser aplicadas no contexto dos dados abertos do Recife. Na sequência, o trabalho descreve os procedimentos metodológicos adotados, desde a coleta e pré-processamento dos dados até a implementação do algoritmo Apriori e o processo de mineração. Por fim, são apresentados os resultados experimentais, com a identificação e análise dos padrões descobertos, evidenciando a aplicabilidade prática das técnicas exploradas.

2 REFERENCIAL TEÓRICO

2.1 Padrões

Nesta seção, o conceito de padrão é revisto sob a perspectiva de diferentes disciplinas. Em mineração de dados, o termo padrão assume um significado operacional e quantitativo. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 83), padrões são "estruturas ou relações identificáveis nos dados que são válidas, novas, potencialmente úteis e compreensíveis". Essa definição estabelece os quatro critérios fundamentais para que uma descoberta em dados seja considerada um padrão: validade estatística, novidade, utilidade e inteligibilidade.

Segundo Han, Kamber e Pei (2011), padrão é uma expressão ou conjunto de características que ocorre com frequência suficiente nos dados para ser considerada relevante. Aggarwal (2015) define padrões como entidades matemáticas ou estatísticas que representam conhecimento latente nos dados, ou seja, que estão ocultos. Esses padrões são extraídos de grandes volumes de dados por meio de algoritmos mineração. Witten, Frank e Hall (2016) enfatizam a importância dos padrões como relações consistentes e interpretáveis, representadas por modelos como regras de classificação ou árvores de decisão.

Nesta área de mineração de dados, padrões são relacionados ao conhecimento por meio de sua capacidade de gerar novas ideias, orientar a decisão e alimentar modelos preditivos. De maneira geral, devem ser úteis para determinados contextos. São expressões quantitativas de regularidades que, uma vez interpretadas, se tornam conhecimento acionável (Fayyad *et al*., 1996).

Nas Ciências Sociais, o conceito de padrões é abordado sob diversas perspectivas teóricas que os reconhecem como construções sociais historicamente situadas. Elias (1939), aborda o conceito de padrões como produtos históricos de processos civilizatórios e interdependências sociais. Garfinkel (1967) discute como as pessoas criam e dão sentido às suas ações cotidianas. Ele destaca que os padrões são formas práticas e compartilhadas que permitem às pessoas interpretar o mundo ao seu redor por meio de interações diárias e saberes comuns. Esses padrões são, portanto, construídos e mantidos pelas práticas sociais.

Do ponto de vista interpretativista, Geertz (1973) concebe padrões culturais como teias de significados compartilhados, nos quais a cultura é composta por símbolos que organizam a experiência humana, enfatizando o papel interpretativo na compreensão das práticas sociais.

Diante desse contexto, P. Bourdieu (1980) por sua vez associa os padrões ao conceito de *habitus*, que se refere a esquemas internalizados de percepção e disposição que moldam as

ações dos indivíduos de forma prática e inconsciente. Esses padrões são criados por meio da interação entre estruturas sociais e experiências individuais ao longo do tempo, sendo transferíveis entre diferentes contextos, mas também sujeitos a mudanças históricas.

Dessa forma, padrões são compreendidos como regularidades significativas que emergem das práticas, relações e estruturas sociais (Giddens, 1984). Essas regularidades são apenas observadas empiricamente, possuindo significado simbólico, histórico e cultural. Para Giddens (1984), os padrões são como elementos constituintes das práticas sociais rotineiras, reproduzidos no processo de estruturação.

Diante dessas contribuições, nas Ciências Sociais, os padrões são inseparáveis do contexto, da interpretação e do conhecimento incorporado (intrínseco) no âmbito social. Mais do que estruturas observadas, são estruturas vividas.

Por fim, tanto nas Ciências Sociais quanto na mineração de dados, padrões representam uma forma de sistematização do conhecimento. Seja por meio da experiência acumulada e do simbolismo cultural como em Bourdieu (1980) ou Geertz (1973), seja via modelagem e abstração computacional como em Han *et al.* (2011), Fayyad *et al.* (1996) ou Aggarwal (2015), os padrões funcionam como pontes entre o mundo empírico e a inteligência estruturada sobre ele.

Nesse sentido, é pertinente resgatar o pensamento de Alexander (1977), cuja obra *A Pattern Language*, define padrões como soluções recorrentes para problemas em contextos complexos. Para Alexander, os padrões têm um papel estruturante: eles conectam regularidades observadas com formas compreensíveis de ação, sejam arquitetônicas, sociais ou de projeto. Embora sua abordagem tenha emergido na arquitetura, ela inspira abordagens computacionais e analíticas.

A partir desse contexto, os padrões têm valor porque explicam fenômenos, revelam estruturas e informam ações. A diferença é que, em *knowledge discovery in database* (KDD), o padrão é muitas vezes descoberto de forma empírica, por métodos automatizados, enquanto para Alexander (1977) e para o campo das Ciências Sociais, ele pode ser formulado a partir da experiência humana vivida.

No âmbito da Ciência da Informação, o conhecimento também é compreendido como uma construção dinâmica e relacional entre sujeito, informação e contexto. Capurro e Hjorland (2003) apontam que o conhecimento é produzido por meio da mediação entre sujeitos e informação, sendo estruturado pelas práticas culturais, sociais e profissionais. Desse modo, padrões só se tornam conhecimento quando fazem sentido para um contexto prático ou intelectual.

Em Gestão do Conhecimento, Nonaka e Takeuchi (1997) interpretam os padrões como mecanismos de externalização do conhecimento tácito, possibilitando que formas de conhecimento não formalizadas, comumente oriundas da experiência e da prática, sejam articuladas e convertidas em conhecimento explícito, compartilhável e reaplicável em diferentes contextos. Isso é visível tanto em linguagens de padrões, como a de Alexander (1977), quanto nos modelos preditivos de mineração de dados. A transformação de conhecimento tácito em explícito por meio de padrões é, portanto, um elo fundamental entre observação empírica, interpretação e ação informada. Assim, o padrão pode ser entendido como uma informação que, quando interpretada, é transformada em conhecimento socialmente construído.

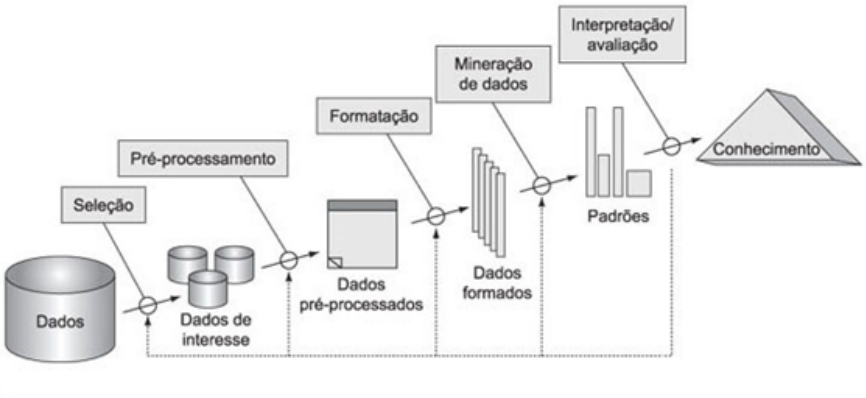
2.2 Descoberta de conhecimento em base de dados

Desde o pós-guerra, vem se reconhecendo, paulatinamente, que a produtividade e a competitividade dos agentes econômicos dependem cada vez mais da capacidade de lidar eficazmente com a informação para transformá-la em conhecimento (Lastres e Albagli, 2002). Essa transformação é essencial em um mundo onde a informação se torna um recurso estratégico.

Nesse contexto, o processo de descoberta do conhecimento em bancos de dados, conhecido como KDD, surge como uma abordagem importante. Este campo, que está em constante evolução, envolve interações com áreas como estatística, aprendizado de máquina e gerenciamento de dados. O KDD tem evoluído significativamente nos últimos 30 anos, desde sua formalização em 1989, em resposta aos desafios enfrentados pelas organizações em lidar com o grande volume de dados armazenados. Nessa época, as técnicas e análises disponíveis não eram suficientes para identificar padrões e associações relevantes dentro de grandes conjuntos de dados, devido a limitações computacionais que impactavam diretamente a eficiência e eficácia dos processos de extração do conhecimento.

De acordo com Fayyad et al. (1996), o processo de KDD envolve uma série de tarefas compostas por etapas para obter informações dos dados e transformá-los em conhecimento útil. Essas relações podem ser melhor visualizadas na Figura 1:

Figura 1 - Processo de descoberta de conhecimento em bases de dados



Fonte: Fayyad; Piatetsky-Shapiro; Smyth, (1996).

O processo de KDD envolve várias etapas e decisões tomadas pelo usuário, essenciais para orientar toda a ação. Dessa forma, é de suma importância compreender o contexto do negócio no qual o conjunto de dados está inserido, para garantir uma análise bem-sucedida. O processo é dado de forma iterativa, o que significa que os dados são continuamente analisados e reavaliados para encontrar as informações necessárias (Fayyad *et al.* 1996). O processo é formado por quatro principais etapas: seleção, pré-processamento, mineração de dados e pós-processamento.

2.2.1 Seleção

Segundo Fayyad et al. (1996), a seleção representa a primeira etapa desse processo, que consiste na atividade de identificar e selecionar informações relevantes (*target data*) para o contexto e objetivo da análise. Essa fase funciona como um filtro informacional, voltado à exclusão de dados redundantes, como atributos e registros irrelevantes ou ruidosos, de modo a otimizar o desempenho das etapas posteriores da mineração.

Goldschmidt e Passos (2006) observam que essa etapa pode envolver tanto a redução horizontal (filtragem de registros) quanto a redução vertical (seleção de atributos), a depender da complexidade e do foco da análise. A seleção pode ser realizada de forma manual, com apoio de especialistas do domínio, ou de maneira automatizada, por meio de baseados em regras ou critérios estatísticos, como amostragem aleatória, simples com ou sem reposição (Goldschmidt e Passos, 2006).

2.2.2 Pré-processamento dos dados

O pré-processamento dos dados consiste em duas etapas para preparar os dados para a mineração, que são detalhados a seguir:

I. Limpeza de dados é o processo de identificar e corrigir ou remover erros, inconsistências e informações irrelevantes em um conjunto de dados. Isso inclui tratar valores ausentes, duplicados, erros de formatação, outliers e dados inválidos. O objetivo é garantir que os dados estejam precisos, consistentes e prontos para análise ou modelagem, melhorando a qualidade dos resultados obtidos.

II. Transformação de dados consiste em ações como a adição de novas colunas, a realização de cálculos sobre colunas já existentes, o agrupamento de variáveis contínuas em faixas ou a conversão de variáveis categóricas em variáveis binárias. Para Edelstein (1998), um dos principais desafios é lidar com as transformações necessárias para gerar previsões, sendo que muitas ferramentas de mineração de dados deixam essa tarefa como responsabilidade do usuário ou programador. Portanto, esse processo consiste na consolidação e transformação dos dados em formas apropriadas para a mineração, especificamente para aplicação do algoritmo, podendo ser realizada por uma série de operações como categorização, transformação de valores em intervalos, conversão de linhas em colunas como também a transformação de dados categóricos para numéricos.

2.2.3 Mineração de dados

A mineração de dados consiste em métodos para extrair padrões ou construir modelos a partir dos dados (Han e Kamber, 2001). Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 83) definem esta etapa como "a aplicação de algoritmos específicos para extrair padrões dos dados". Os dois principais objetivos na mineração de dados, são verificação e descoberta. A verificação envolve corroborar uma hipótese do pesquisador, enquanto a descoberta envolve a atividade de encontrar novos padrões.

Para cada problema de mineração de dados, existem algoritmos adequados para obter uma solução satisfatória. Esses algoritmos podem executar dois tipos de tarefas, apresentados a seguir:

I. Tarefas descritivas: concentram-se em encontrar padrões que descrevam os dados de forma interpretável pelos seres humanos. Agrawal *et al.* (1993) definem a extração de regras de associação e o agrupamento (*clustering*) como as principais tarefas descritivas, na qual a extração de regras de associação visa identificar relações frequentes entre itens em grandes conjuntos de dados, enquanto o agrupamento organiza os dados em grupos homogêneos com base em características semelhantes.

II. Tarefas preditivas: concentram-se em inferir informações sobre os dados existentes para prever o comportamento de novos dados. De acordo com Han e Kamber (2001), as principais tarefas preditivas incluem a classificação e a regressão.

2.2.3.1 Regras de associação

O acúmulo massivo de dados gerados tem impulsionado a necessidade de técnicas e métodos eficientes para extrair conhecimento útil a partir de grandes volumes de informações. Nesse contexto, a pesquisa realizada por Agrawal *et al.* (1993) destaca que a evolução da tecnologia de código de barras facilitou a coleta detalhada de dados transacionais, como compras no varejo, sendo a 'análise de cestas de mercado' (*basket data*) um exemplo clássico dessa abordagem.

Agrawal *et al.* (1993) desenvolveram um estudo fundamental sobre a descoberta de regras de associação em grandes bases de dados transacionais. Seu trabalho explora como essas técnicas podem ser aplicadas a partir da análise de dados para identificar tendências ou padrões no comportamento de compra dos consumidores. Esse processo permite revelar associações interessantes entre itens, úteis para previsão e tomada de decisão (Tsay e Chiang, 2005).

Para compreender o conceito de regras de associação, é importante primeiro entender o significado de "regra" e "associação". De acordo com Agrawal *et al.* (1993) uma regra pode ser definida como um princípio que estabelece uma relação lógica entre elementos distintos. Já uma associação refere-se à ligação ou correlação entre diferentes elementos que compartilham alguma relação de interdependência. No contexto da mineração de dados, as regras de associação representam relações probabilísticas entre conjuntos de itens em uma base transacional, indicando a chance de certos itens serem adquiridos juntos, seguindo o conectivo lógico direcional, no qual "Se X, então Y".

Também é necessário compreender também os conceitos de bases transacionais e conjuntos de itens. Segundo Agrawal *et al.* (1993), uma base transacional é um tipo específico de banco de dados *D* que armazena registros de transações realizadas por usuários, clientes ou sistemas. Cada transação *T* é identificada por um código único *TID* e representa um conjunto de itens adquiridos ou registrados em uma única operação, denominado *itemset*. Por exemplo, um supermercado pode ter uma base de dados que registra todas as compras feitas pelos clientes, onde cada transação corresponde a uma compra completa, compreendendo todos os produtos adquiridos.

Pode-se representar uma base transacional conforme a Tabela 1 a seguir:

Tabela 1 – Exemplo de estrutura de uma tabela de transações.

id_transacao (TID)	itemset
t1	pão, leite, manteiga
t2	leite, queijo, manteiga
t3	pão, café
t4	leite, manteiga, café

Fonte: Elaborado pelo autor, 2025.

Suponha que um cliente, em uma transação, compra pão e manteiga, então existe alta probabilidade de comprar leite. Essa relação pode ser expressa por meio da seguinte regra de associação: {pão, manteiga}⇒{leite}. O conjunto {pão, manteiga} é denominado antecedente da regra e o conjunto {leite} é o consequente, que representa o item que tem uma probabilidade de ser adquirido quando o antecedente ocorre. Essa relação indica que, dentro da base de dados analisada, sempre que os itens pão e manteiga aparecem juntos em uma compra, há uma probabilidade de que o leite também seja adquirido.

De acordo com Agrawal *et al.,* (1993) regras de associação podem ser formalizadas da seguinte forma: $X \Rightarrow Y$, onde X e Y são conjuntos de itens pertencentes a um conjunto de dados I, garantindo que X e Y sejam subconjuntos distintos ($X, Y \subseteq I$) e que não possuam elementos em comum ($X \cap Y = \emptyset$). Segundo Agrawal e Srikant, 1997, p. 3) "uma regra de associação generalizada é uma implicação da forma $X \Rightarrow Y$, onde $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \emptyset$, e nenhum item em Y é ancestral de qualquer item em X". Dessa forma, a regra expressa a relação de que, se um conjunto de itens X aparece em uma transação, há uma probabilidade de que o conjunto de itens Y também esteja presente.

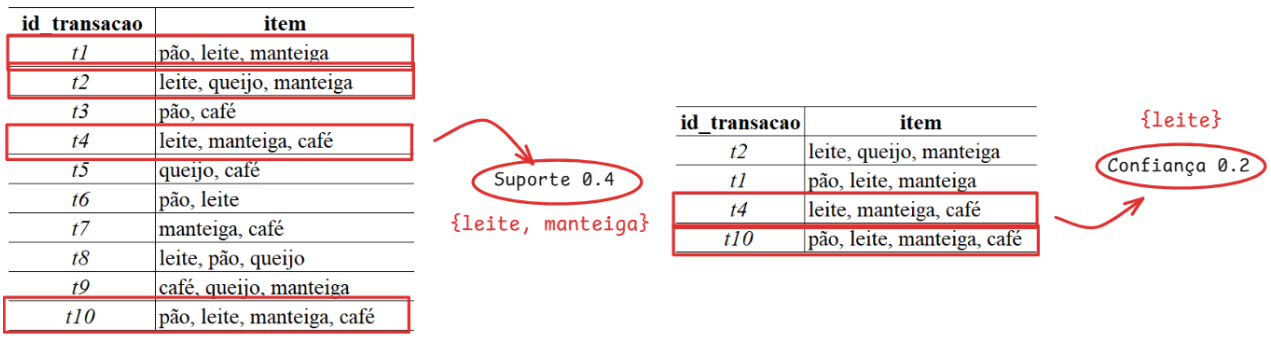
Essa relação estatística é quantificada por duas métricas apresentadas a seguir:

I. Suporte (*s*) é uma medida que indica a fração de transações com que um conjunto de itens X e Y ocorre em todo o conjunto de dados. Ele é calculado como a proporção do número de transações que contêm o conjunto de itens em relação ao número total de transações no banco de dados. O suporte funciona como um filtro para determinar quais conjuntos são frequentes o suficiente para serem considerados relevantes.
II. Confiança (*c*) mede a probabilidade condicional da presença do consequente Y dado a presença do antecedente X em uma regra associativa. Em outras palavras, ela expressa o quanto provável é que os itens subsequentes ocorram quando os anteriores estão pre-

sentes. A confiança ajuda na identificação das relações mais significativas entre os conjuntos. Ela garante que apenas as associações fortes sejam mantidas após aplicada um limiar mínimo.

Assim, pode-se sintetizar todos esses elementos a partir da Figura 2:

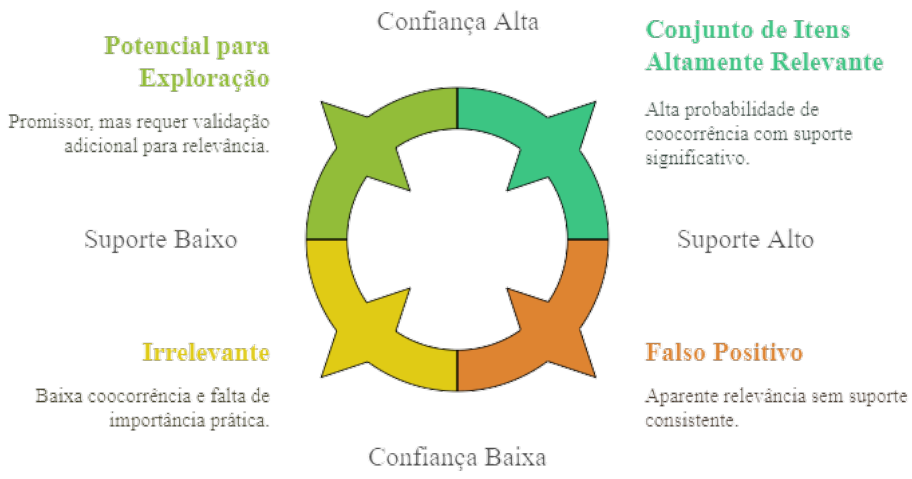
Figura 2 - Suporte e confiança



Fonte: Elaborado pelo autor, 2025.

Essas métricas são necessárias definir critérios e avaliar a relevância e a utilidade das regras extraídas, garantindo que as saídas (*output*) tenham aplicabilidade prática com o objetivo da mineração. Ao analisar regras de associação, torna-se importante compreender como suporte e confiança interagem para definir a relevância dos padrões descobertos. A Figura 3, nesse sentido, sintetiza essa relação entre suporte e confiança, facilitando a interpretação dos resultados e fornecendo um guia visual para classificar as regras de associação descobertas, de acordo com os valores dessas métricas:

Figura 3 - Classificação de regras de associação com base em suporte e confiança



Fonte: Elaborado pelo autor, 2025.

2.2.3.2 Algoritmo Apriori

A seção anterior abordou o conceito de regras de associação, com destaque para sua importância na mineração de dados para identificar padrões frequentes dentro de bases transacionais. Contudo, para que essas regras possam ser extraídas de forma eficiente, é necessário o uso de algoritmos sofisticados, capazes de lidar com grandes volumes de dados.

No estudo inicial de Agrawal *et al.* (1993), os autores introduziram o algoritmo AIS (*Artificial Intelligence System*) para a descoberta de regras de associação. Esse algoritmo permitia a identificação de padrões frequentes a partir de um banco de dados transacional. Contudo, apresentava limitações significativas, como o alto custo computacional uma vez que para identificar combinações de *itemsets* é necessário percorrer repetidamente a base de dados.

Diante dessas limitações, Agrawal e Srikant (1994) propuseram uma abordagem aprimorada, resultando no desenvolvimento do algoritmo Apriori, que estabelece que um subconjunto de um *itemset* frequente também deve ser frequente. Isso permitiu reduzir drasticamente o número de combinações analisadas, eliminando candidatos infrequentes já nas primeiras iterações. Assim, a mineração de regras de associação tornou-se mais eficiente e escalável.

Segundo Zaki *et al.* (1997), o Apriori foi um avanço significativo para a descoberta de regras de associação, sendo amplamente aceito como um dos algoritmos mais eficazes. O algoritmo pode ser dividido em duas etapas:

I. Descoberta de conjuntos frequentes: nesta fase, são identificados os conjuntos de itens que ocorrem com uma frequência mínima especificada pelo usuário. O Apriori aplica múltiplas varreduras no banco de dados para refinar os candidatos.
II. Geração de regras de associação: com os conjuntos frequentes identificados, são geradas regras na forma $X \Rightarrow Y$, garantindo que a relação entre os itens seja estatisticamente significativa.

No entanto, mesmo com o Apriori, o processo de mineração de regras de associação ainda apresentava desafios computacionais, como o número excessivo de varreduras no banco de dados, que impactava a eficiência em bases de dados massivas. Zaki *et al*. (1997) aponta que, para bases de dados grandes, novas abordagens poderiam ser adotadas para minimizar os custos computacionais.

A descoberta de regras de associação é um dos principais desafios da mineração de dados, sendo essencial otimizar o processo para encontrar os conjuntos de itens mais relevantes dentro

de grandes bases de dados. Além disso, foram propostos novos algoritmos que buscam melhorar o desempenho do Apriori e reduzir o número de leituras do banco de dados. Segundo Zaki *et al.* 1997*,* o Apriori foi demonstrado superior às abordagens anteriores (Park *et al.* 1995); Holsheimer *et al.* 1995). A vantagem do Apriori sobre abordagens anteriores, é que ele reduz o espaço de busca através da propriedade de fechamento descendente, ou seja, eliminando candidatos infrequentes antes da execução completa do algoritmo.

Contudo, estudos posteriores sugerem que novas abordagens, como o algoritmo FP-Growth e métodos baseados em *clustering* e taxonomia de *itemsets*, podem oferecer ganhos ainda maiores na descoberta de regras de associação sem a necessidade de múltiplas leituras do banco de dados.

2.2.4 Pós-processamento

Nesta etapa, ocorre a interpretação dos padrões descobertos que são interessantes e úteis com base em critérios de relevância (Han e Kamber, 2001). Os critérios de relevância referem-se às métricas e métodos utilizados para determinar a importância e utilidade dos padrões ou modelos descobertos. São avaliados também critérios subjetivos, baseados no conhecimento de especialistas e na viabilidade de aplicação dos padrões descobertos (Goldschmidt e Passos, 2006). Um padrão pode ser estatisticamente relevante, mas irrelevante do ponto de vista estratégico ou operacional, o que exige julgamento humano para sua validação final. Outro aspecto importante dessa fase é a visualização dos padrões, que contribui para tornar os resultados mais acessíveis e interpretáveis.

Além disso, a avaliação pode apontar necessidades de retorno a etapas anteriores, como ajustes no pré-processamento, na seleção de atributos ou até na própria definição do problema. Isso reforça a natureza iterativa do processo de KDD, onde cada etapa é interdependente e passível de revisão com base nos resultados alcançados.

Assim, a descoberta de conhecimento não se resume à aplicação mecânica de algoritmos, mas a um processo reflexivo, em que a compreensão e o julgamento humano desempenham papel relevante.

2.3 Enriquecimento de dados

Para Shyalika *et al.* (2024), o enriquecimento de dados se compõe de três principais técnicas: aumento de dados, amostragem e imputação, cada qual desempenhando um papel especí-

fico na melhoria dos conjuntos de dados. O aumento de dados, amplamente utilizado em aprendizado de máquina, tem como objetivo expandir o tamanho e as características dos conjuntos de dados, garantindo maior representatividade das bases disponíveis. Essa técnica envolve a criação de novos atributos a partir de transformações nos dados originais, enriquecendo a diversidade e complexidade das informações analisadas. Dessa forma, a técnica de aumento de dados pode ser entendida como o processo de adicionar novas características a uma variável ou a um conjunto de variáveis, com o objetivo de ampliar seus atributos e facilitar interpretações mais profundas. Ao incorporar informações externas, os dados ganham um nível de compreensão mais detalhado.

A técnica de amostragem (*sampling*) é usada para equilibrar conjuntos de dados desbalanceados, no qual uma classe (como eventos raros) tem muito menos exemplos do que outra classe mais comum. Isso é importante porque modelos de aprendizado de máquina tendem a favorecer a classe maior.

Shyalika *et al.,* (2024) destacam duas abordagens principais: *oversampling*, que aumenta o número de amostras da classe minoritária por meio da replicação ou criação de novas amostras, e *undersampling*, que reduz o número de amostras da classe majoritária removendo exemplos redundantes. O resultado é um conjunto de dados equilibrado, com proporções similares entre as classes, permitindo que os modelos sejam mais justos e precisos na identificação das duas categorias. Por fim, a técnica de imputação aborda a correção de valores incorretos e o preenchimento de registros nulos na base de dados, contribuindo para a consistência e integridade das informações analisadas.

Quando se adiciona informações complementares a uma base de dados, tornam os modelos mais robustos e preparados para identificar padrões e tomar decisões. Dessa forma, o enriquecimento de dados funciona como uma ferramenta que expande as possibilidades, permitindo que análises sejam mais completas e precisas. O enriquecimento de dados beneficia os modelos de aprendizado de máquina, conferindo-lhes capacidades aprimoradas para compreender padrões complexos, generalizar efetivamente e fazer previsões mais precisas.

Entretanto, essa expansão pode ser feita de forma mais racional e estratégica com o objetivo de reduzir esforços desnecessários e desperdício de recursos. Em outras palavras, ao planejar cuidadosamente quais informações adicionais serão incorporadas, pode-se otimizar o processo e evitar trabalhos desnecessários.

Esses esforços, no contexto da análise de dados, podem ser associados ao tempo de processamento do algoritmo utilizado na mineração. Ao enriquecer os dados de forma direcionada e progressiva, pode-se garantir não apenas uma maior representatividade para entendimento das

saídas *(outputs)* e qualidade das informações, mas também um uso mais eficiente dos recursos computacionais, o que torna todo o processo mais ágil e eficiente. Isso ocorre porque, ao incorporar apenas as informações relevantes e estrategicamente selecionadas, reduzimos a complexidade e o volume de dados desnecessários, diminuindo o tempo de processamento e evitando sobrecarga nos algoritmos. De acordo com Han, Kamber e Pei (2011), medidas de interesse de padrões, objetivas ou subjetivas, podem ser usadas para orientar o processo de descoberta. Assim, os esforços se concentram em informações que realmente contribuem para o objetivo final, otimizando o fluxo de trabalho e os resultados obtidos.

Como o processo de mineração de dados é iterativo, é necessário frequentemente retornar ao estado inicial da análise para executar novamente os algoritmos, ajustando os parâmetros e refinando os dados a cada etapa. Quando se utiliza algoritmos muito robustos ou complexos, o tempo de processamento pode aumentar significativamente, dependendo do modelo e do tamanho dos dados aplicados na análise. Esse fator torna ainda mais essencial a racionalização e o direcionamento do enriquecimento de dados, garantindo que o ciclo iterativo seja eficiente e sustentável em termos de recursos computacionais.

A próxima seção apresenta os procedimentos metodológicos adotados nesta pesquisa, descrevendo as abordagens, ferramentas e critérios utilizados para a condução da descoberta.

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa adota uma abordagem quali-quantitativa, utilizando tanto métodos quantitativos para a análise de dados estruturados quanto métodos qualitativos para a interpretação dos resultados obtidos (Knechtel, 2014). O objetivo desta pesquisa é descobrir e analisar padrões e associações presentes nas infrações de trânsito ocorridas na cidade do Recife. A partir dessa análise, busca-se correlacionar tais eventos com variáveis temporais, de modo a interpretar o comportamento dessas infrações ao longo do tempo. Essa abordagem permite delinear os fins do estudo, que se concentram na descrição detalhada desses fenômenos e na exploração de possíveis relações que contribuam para o entendimento mais amplo da dinâmica do trânsito urbano.

De acordo com os fins, a pesquisa se configura como exploratória e descritiva, em conformidade com os objetivos e métodos adotados ao longo do estudo. Classifica-se exploratório, uma vez que se propõe a investigar os dados, empregando técnicas de mineração de dados com o intuito de identificar padrões ocultos e levantar hipóteses iniciais sobre o comportamento urbano (Hair, 2005). Também possui caráter descritivo, pois visa apresentar e detalhar fenômenos específicos relacionados às infrações de trânsito registradas na cidade do Recife (Koche, 2012).

De acordo com os meios, a pesquisa pode ser classificada como bibliográfica pois se apoia em autores clássicos e contemporâneos para fundamentar suas análises e interpretações. A revisão da literatura teve papel essencial na definição do problema, na delimitação do objeto e na interpretação dos resultados Lakatos e Marconi (2003), e documental, uma vez que utiliza bases de dados públicas que possui caráter de fontes primárias, disponibilizadas por instituições oficiais, como o Portal de Dados Abertos do Recife disponibilizada pela prefeitura em parceria com a Empresa Municipal de Informática (EMPREL).

3.1 Coleta de dados

A obtenção de dados brutos originários de diferentes fontes representa o início do processo metodológico. Essa etapa foi fundamental para reunir informações que possibilitaram posteriormente a aplicação do algoritmo adotado para pesquisa e análise detalhada permitindo a identificação de padrões e associações relevantes entre as bases. Os dados foram extraídos de duas fontes primárias heterogêneas: o Portal de Dados Abertos do Recife.

O Portal forneceu uma base estruturada das infrações de trânsito cometidas na cidade disponíveis no portal como ‘Registro das Infrações de Trânsito’, contendo informações detalhadas,

como tipo de infração cometida, amparo legal, local da infração, descrição, código da infração, data e horários dos cometimentos e agente equipamento.

As bases foram coletadas a partir do dia 01 de janeiro até o dia 31 de dezembro, nos anos de 2020 a 2023. Para cada ano foi realizada uma exportação de tabela, totalizando 4 exportações diferentes no portal da prefeitura. Tais documentos também estavam disponíveis no formato CSV (*Comma-Separated Values*), amplamente utilizado para manipulação e disseminação de grande volume de dados. Os dados foram fornecidos sob a Licença Aberta para Bases de Dados (ODbL), que permite o uso, modificação e compartilhamento dos dados, desde que mantida a mesma licença em trabalhos derivados.

3.2 Pré-processamento

Os dados, após serem coletados, passaram por um processo de transformação e modelagem, seguido pelo enriquecimento dos dados, como detalhados a seguir.

3.3.1 Transformação dos dados

Os dados coletados foram ajustados para garantir compatibilidade, permitindo posteriormente a aplicação do algoritmo Apriori. Para o tratamento e organização dos dados, utilizou-se o software Microsoft Excel (Versão 2502 Build 16.0.18526.20168) 64 bits, seguindo um procedimento metodológico estruturado em múltiplas etapas. O primeiro passo consistiu na importação dos dados da fonte original para o Excel, onde foi aplicada a técnica de pivotagem (*unpivot*) para reorganizar a estrutura da tabela, tornando-a mais adequada para análise e manipulação futura.

Na etapa de tratamento e organização dos dados referentes às infrações, efetuaram-se os seguintes passos. O primeiro passo consistiu na exportação dos arquivos de dados sobre as multas para o software Excel diretamente do portal de dados disponibilizado em: Organizações → Autarquia de Trânsito e Transporte Urbano do Recife – CTTU → Registro das Infrações de Trânsito. Após a importação dos dados para o Excel, foi realizada a formatação dos arquivos para o padrão UTF-8. Essa etapa foi crucial, pois garantiu que todos os caracteres fossem lidos corretamente durante a importação para o banco de dados. A formatação UTF-8 assegura que informações como acentuação, caracteres especiais e símbolos sejam preservados, evitando possíveis erros de leitura ou perda de dados.

Após todas as bases serem modeladas e formatadas utilizando o Excel, elas estavam prontas para leitura e importação para o Sistema de Gerenciamento de Banco de Dados (SGBD)

PostgreSQL na versão 13. No SGBD, foram seguidos passos rigorosos que garantiram a integridade e a organização das informações. Para isso, criou-se um esquema para a fonte de dados, nomeado de ‘recife_multas’. Essa estruturação permite uma gestão mais eficiente e clara dos dados, facilitando a construção de futuras consultas SQL, análises e manutenções.

Após a criação dos esquemas no banco de dados, foram estabelecidas tabelas específicas para cada ano, assegurando que os dados fossem armazenados de maneira organizada e cronológica. Cada tabela foi configurada com base nas melhores práticas de modelagem de banco de dados, incluindo o preenchimento de metadados que documentam as características das colunas e suas finalidades. Além disso, foram aplicadas restrições essenciais para garantir a integridade dos dados, como:

a) *PRIMARY KEY*: Cada tabela recebeu uma chave primária para garantir um identificador único por registro. Na base de infrações, por exemplo, a coluna "id_multas", criada com a função *SERIAL*, foi definida como chave primária.
b) *UNIQUE*: Para garantir que não haja valores duplicados em colunas específicas.
c) *NOT NULL*: Para assegurar que determinadas colunas não aceitem valores nulos, garantindo a presença obrigatória de informações.
d) *FOREIGN KEY*: Para estabelecer relacionamentos entre tabelas e garantir a integridade referencial.

Diante desses recursos, a importação foi realizada utilizando comandos apropriados no PostgreSQL, garantindo que todos os dados fossem introduzidos corretamente nas tabelas designadas. Esse processo estruturado assegurou que as informações estivessem organizadas e prontas para as devidas manipulações.

Após o carregamento das tabelas no (SGBD), procedeu-se à implementação do algoritmo Apriori. Paralelamente, foram aplicadas técnicas de enriquecimento dos dados, fundamentais para ampliar a capacidade analítica do algoritmo. Essas transformações consistiram na criação de atributos temporais derivados das variáveis de data, os quais foram correlacionados com outras colunas do conjunto de dados, como o tipo de infração cometida. Tal abordagem visou não apenas refinar os resultados gerados pelo algoritmo, mas também aprofundar a compreensão das relações contextuais e temporais entre os eventos analisados.

3.3.2 Implementação do algoritmo Apriori

O algoritmo Apriori foi codificado em linguagem SQL e implementado no software Agadê Mineração, versão 4.6, disponível em https://agade.ufpe.br/, que está instalado no Postgresql 13 do servidor do Laboratório Agadê.

Para o processo de preenchimento das tabelas utilizadas pelo algoritmo Apriori, foi adotada uma estrutura organizada de enriquecimento de dados. Esse processo teve como objetivo transformar atributos brutos em características mais relevantes e significativas para a etapa de mineração de dados.

O algoritmo implementado usa duas tabelas, que são: 'apriori_itens', que cria um vocabulário de itens distintos e; a associação desses itens a transações na tabela 'apriori_transacoes', que permite o estabelecimento de relações entre os eventos analisados. Para isso, foram atribuídos identificadores únicos a cada item, permitindo que, em cada transação registrada, os atributos fossem representados de maneira padronizada e estruturada. Composta pelas seguintes estruturas:

I. Tabela 'apriori_itens': Responsável pelo armazenamento de todos os atributos distintos encontrados nas transações da base analisada. Cada item recebe um identificador único 'id_item' e é associado a um identificador de atributo 'id_atributo', que representa a categoria a qual pertence. Além disso, o campo 'nome', armazena a descrição textual do item, facilitando a interpretação dos padrões identificados. Dentro desse contexto:
    a. id_item: Representa um valor único existente dentro de uma categoria. Exemplo: "manhã" como um valor dentro do atributo 'Período do Dia'.
    b. id_atributo: Indica a categoria ou característica geral a que um item pertence Exemplo: 'Período do Dia'
    c. nome: Contém a descrição completa do item, permitindo que as análises resultantes sejam mais compreensíveis.
II. Tabela 'apriori_transacoes': Estruturada com três colunas: 'id_transacao', 'id_item' e 'id_atributo'. A coluna 'id_transacao' representa cada registro único na base original, enquanto 'id_item' e 'id_atributo' estabelecem a relação entre os itens e seus atributos correspondentes. Dessa forma, cada transação armazena um conjunto de itens associados a diferentes atributos, permitindo a análise de padrões recorrentes.
III. Tabela 'apriori_regras': Utilizada para armazenar as regras de associação geradas pelo algoritmo Apriori após a mineração dos dados. Essa tabela é essencial para consolidar

os padrões identificados e permitir a interpretação das correlações encontradas. As colunas dessa tabela incluem:

a. id_regra: Identificador único de cada regra de associação gerada.
b. antecedentes: Conjunto de itens que compõem a parte inicial da regra. Esses itens são aqueles que, quando presentes, podem levar à ocorrência dos consequentes.
c. consequentes: Conjunto de itens que compõem a parte final da regra (consequentes), ou seja, os itens que têm maior probabilidade de ocorrer quando os antecedentes estão presentes.
d. nomes_antecedentes: Descrição textual dos antecedentes, facilitando a interpretação dos resultados.
e. nomes_consequentes: Descrição textual dos consequentes.
f. k: Número de itens combinados na regra, indicando o tamanho da associação.
g. confianca: Métrica que indica a confiabilidade da regra gerada, representando a proporção de transações em que os antecedentes levam aos consequentes.
h. suporte: Frequência relativa da ocorrência da regra no conjunto de dados analisado, indicando a relevância estatística do padrão encontrado.
i. registros: Número total de transações que contêm tanto os antecedentes quanto os consequentes.

O modelo lógico das tabelas é apresentado na Figura 4:

Figura 4 - Modelo lógico das tabelas do algoritmo Apriori

Fonte: Elaborado pelo autor, 2025.

Para ilustrar a aplicação do modelo lógico do algoritmo Apriori utilizado, foi realizada a população das tabelas com um conjunto de transações simuladas. Os itens foram registrados com seus respectivos identificadores a partir da coluna id_atributo, onde os atributos podem ter 1 ou mais itens como apresentado na tabela a seguir com os atributos ‘03 – Hora’, ‘10 – ano’ e ‘0 – multa’:

Tabela 2 - Representação da tabela apriori_itens populada

id_atributo	*id_item*	*nome*
00	1	Multas
03	6	Horário: 06
03	13	Horário: 13
03	18	Horário: 18
03	21	Horário: 21
10	1	Ano: 2020
10	2	Ano: 2021
10	3	Ano: 2022
10	4	Ano: 2023

Fonte: Elaborado pelo autor, 2025.

A tabela ‘apriori_transacoes’, representada na Tabela 3, é responsável por armazenar as transações realizadas, relacionando um identificador de transação id_transacao com os itens adquiridos:

Tabela 3 - Representação da tabela apriori_transacoes populada

id_transacao (TID)	*id_item*	*id_atributo*
t1	1	0
t1	6	3
t1	4	10
t2	1	0
t2	18	3
t2	2	10

Fonte: Elaborado pelo autor, 2025.

Como exemplo, a transação T1 contém os itens ‘multas’, ‘horário: 6’ e ‘ano: 2023’, enquanto a transação T2 contém ‘multas’, ‘horário: 18’ e ‘ano: 2021’. Esses registros demonstram como as transações são armazenadas no banco de dados, garantindo que as regras de associação possam ser extraídas posteriormente. A estrutura permite a aplicação do algoritmo Apriori para identificar conjuntos frequentes de itens, contribuindo para o entendimento do comportamento dos dados.

O processo de transformação e enriquecimento dos dados foi realizado de forma implícita durante as operações de inserção (*INSERT*) nas tabelas específicas destinadas à execução do algoritmo Apriori.

O processo de povoamento inicia-se com a definição do atributo básico de multa, que é identificado e armazenado como um item de referência geral para cada transação. Posteriormente, os atributos passam a ser enriquecidos com informações, derivadas de variáveis temporais presentes na tabela de multas.

O atributo "Data" é gerado a partir da coluna ‘datainfracao’, onde a data é convertida para um formato inteiro padronizado (YYYYMMDD).

A extração do atributo "Dia da Semana" é realizada utilizando a função *EXTRACT(NOW FROM data)* da linguagem SQL, que converte datas em valores numéricos de 0 a 6, representando os dias de domingo a sábado. O atributo resultante recebe uma descrição textual que corresponde ao dia da semana, que permite a identificação de comportamentos de incidência de infrações ao longo da semana.

A segmentação das horas é introduzida com a criação do atributo ‘Hora’, cujos valores que extrai a hora foram extraídos de cada multa. Posteriormente, esse atributo é refinado no conceito de ‘Hora de Pico’, que classifica o período de ocorrência das multas em intervalo de horários de maior fluxo, especificamente das 6h às 9h e das 16h às 19h, durante os dias úteis.

A relação entre a ocorrência de multas e o expediente comercial é definido no atributo ‘Expediente Comercial’. Esse atributo é derivado das informações de data e hora, onde são classificadas como dentro do expediente as infrações ocorridas entre 9h e 18h nos dias úteis.

A identificação de feriados é realizada com apoio de uma tabela de referência com datas predefinidas. Esse enriquecimento permite diferenciar infrações ocorridas em dias regulares e em feriados, auxiliando na identificação de padrões específicos de comportamento viário. Uma abordagem semelhante é utilizada para determinar se a infração ocorreu durante o período de ‘Férias Escolares’, definido pelos meses de janeiro, julho e dezembro.

3.3 Mineração de dados

O processo de mineração de dados foi conduzido de forma iterativa, permitindo ajustes progressivos com base nos resultados obtidos em cada execução. A abordagem adotada buscou otimizar a identificação de padrões relevantes, considerando a dimensão e complexidade da base de dados. Após o enriquecimento, o conjunto de dados passou a contar com 13 atributos distintos, e a tabela ‘apriori_trasacoes’ armazenava, em média, cerca de 7 milhões linhas por ano analisado, tal como exposto na Tabela 4:

Tabela 4 - Atributos enriquecidos

ATRIBUTO	NOME
00	Multas
01	Data
02	Dia da Semana
03	Hora
04	Hora de Pico
05	Expediente Comercial
06	Feriado
07	Férias Escolares
08	Turno do Dia

09	Local do Cometimento
10	Ano
11	Mês
12	Infração

Fonte: Base de dados da pesquisa enriquecida, 2025.

Diante desse cenário, inicialmente, foram empregados critérios mais restritivos, o que resultou em um número reduzido de regras de associação nas primeiras execuções do algoritmo. Na primeira fase da mineração, utilizou-se um suporte mínimo de 0.7 e confiança de 0.8. No entanto, devido ao tamanho da base de dados, essa configuração limitou a descoberta de padrões mais específicos. Além disso, observou-se que a redução dos valores de suporte impactava diretamente no tempo de processamento, que se elevava exponencialmente conforme o suporte mínimo diminuía. Para cada ano analisado, a execução do algoritmo com esses parâmetros iniciais levava, em média, 48 horas para completar, com todos os atributos inseridos na execução.

Diante dessas limitações, foi adotada outra abordagem, com o objetivo de refinar e otimizar a extração das regras relevantes. As modificações foram:

I. Adição de um novo atributo ‘0 multa’ que relaciona todas as ocorrências de infrações na base;
II. Ajuste dos parâmetros de suporte e confiança para 0,2 e 0,8, respectivamente, buscando um equilíbrio entre precisão e volume de regras geradas;
III. Seleção e processamento segmentado de atributos para correlação;
IV. Redução progressiva do suporte conforme a identificação de regras de maior relevância para a análise, atingindo o intervalo de 0,01.

À medida que as regras identificadas se mostravam promissoras, foi possível diminuir ainda mais o suporte mínimo, permitindo a extração de regras mais detalhadas. Na etapa final do processo, com os parâmetros ajustados para sempre considerar ‘multa’ e ‘infração’ como atributos centrais para a análise de correlações com variáveis temporais, como horários, meses e ano. Essa metodologia possibilitou a extração sistemática de regras de associação que evidenciaram relações estatisticamente relevantes entre as variáveis monitoradas.

A mineração de dados foi realizada utilizando o algoritmo Apriori, implementado por meio de procedimentos armazenados codificados em linguagem SQL (*Stored Procedure*). Esse algoritmo foi escolhido por sua capacidade de identificar padrões frequentes em bases de dados

transacionais extensas, garantindo a extração eficiente de associações entre diferentes variáveis. A implementação via linguagem SQL permitiu uma execução otimizada e modular, facilitando a aplicação iterativa do modelo.

3.4 Pós-processamento

Na etapa de pós-processamento, os padrões descobertos foram analisados para determinar se se trata de uma coincidência ou não.

Para isso, foi aplicada uma análise nomológica dos padrões descobertos. A análise nomológica é um método de validação teórica, no qual os padrões identificados são avaliados a partir de construtos teóricos predefinidos, permitindo verificar se as associações encontradas têm coerência dentro de um modelo conceitual estabelecido. Essa abordagem garante que os padrões identificados não sejam apenas associações aleatórias, mas que tenham relevância e embasamento teórico, possibilitando uma interpretação mais robusta e fundamentada dos resultados (Cronbach e Meehl, 1955).

A linguagem de programação Python (Cf. APÊNDICE A) foi empregada para a normalização dos dados e construção de visualizações gráficas, visando a padronização e facilitação da identificação de padrões relevantes durante o processo de mineração de dados.

A normalização dos dados é uma etapa crucial no pré-processamento, que garante que todas as variáveis contribuam igualmente para a análise, evitando vieses decorrentes de diferentes escalas. Conforme destacado por Han et al. (2011), a normalização transforma os dados para um intervalo comum entre 0 e 1. Para isso, foi utilizada a técnica de normalização Min-Max (Ali e Faraj, 2014), que transforma os valores de uma variável para um intervalo entre 0 e 1, conforme a seguinte fórmula:

$$Xnorm = \frac{X - Xmin}{Xmax - Xmin} \quad (1)$$

Em seguida, os resultados serão apresentados de forma estruturada, destacando cada um dos padrões identificados. Cada padrão será detalhado conforme o Quadro 1:

Quadro 1 - Detalhamento dos padrões descobertos

ATRIBUTO DO PADRÃO	DETALHAMENTO
Número	Identificação do padrão encontrado

Título	Título do padrão.
Elementos	Casos em que o padrão foi verificado.
Regra	Descrição da regra de associação correspondente ao padrão.
Descrição	Verificação que os elementos possuem a regra.
Explicação hipotética	Hipóteses sobre a possível explicação do padrão.
Associações	Descrição das associações entre os padrões.

Fonte: Elaborado pelo autor, 2025.

No contexto desta pesquisa, utilizou-se a análise nomológica para investigar as razões pelas quais determinados padrões se repetem, buscando compreender quais variáveis influenciam esses comportamentos e de que maneira essas correlações podem ser explicadas dentro de um modelo conceitual de infrações de trânsito em condições temporais e ambientais. Essa abordagem permitiu explorar fatores subjacentes que podem estar potencializando ou moderando as infrações em dias chuvosos, como:

I. Comportamento dos condutores em condições adversas;
II. Relacionamento das multas em determinados horários do dia.
III. Distribuição temporal das multas.

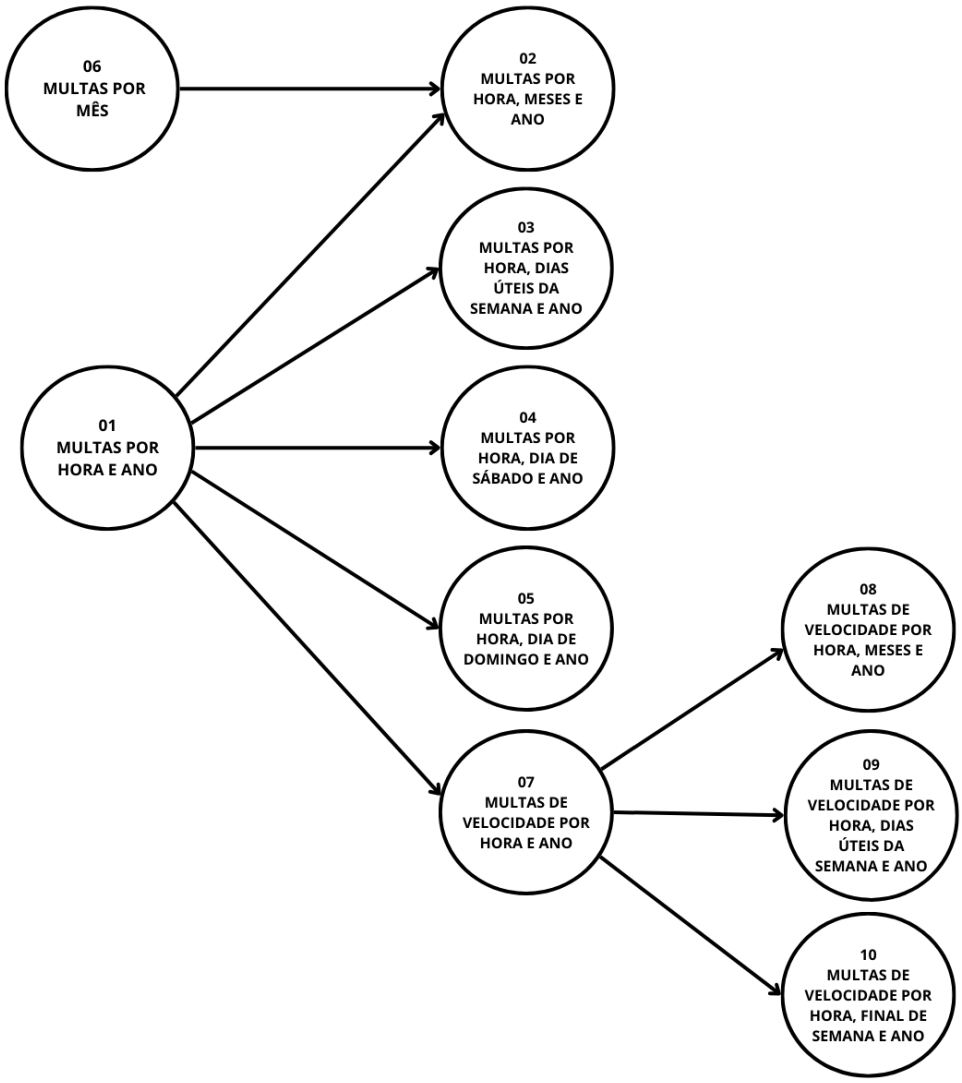
Para essa análise, utilizamos construtos teóricos específicos, que são conceitos estruturados que representam um fenômeno de interesse dentro de um modelo teórico. No caso desta pesquisa, os principais construtos analisados foram os seguintes: 1ª ‘Influência horária no comportamento do condutor’, que engloba a relação entre multas e fatores temporais; 2ª ‘Correlação entre multas e infrações específicas’.

A associação entre os padrões encontrados e os construtos teóricos permite que a análise vá além da simples identificação de correlações, proporcionando uma compreensão mais profunda dos fenômenos observados. Dessa forma, as regras de associação extraídas foram interpretadas dentro de um contexto conceitual, assegurando que os resultados fossem consistentes e aplicáveis a estudos futuros e à formulação de políticas públicas voltadas à gestão do trânsito.

4 RESULTADOS EXPERIMENTAIS

Nesta seção, os padrões descobertos são estruturados e analisados. Identificaram-se 10 padrões que emergiram de forma recorrente nos elementos analisados. Esses padrões foram estruturados em um grafo direcionado sem pesos, que representa as associações entre eles, como apresentado na Figura 6:

Figura 5 - Grafo de associação dos padrões descobertos



Fonte: Elaborado pelo autor, 2025.

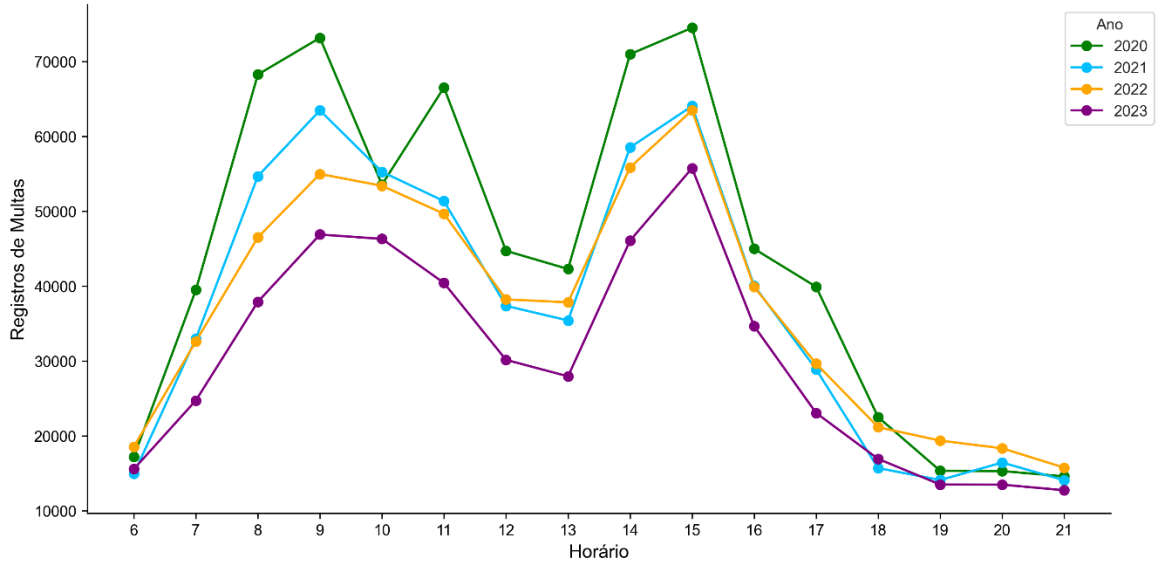
Cada nó representa um padrão e cada aresta representa um aumento de especificidade da origem (mais genérico) para o destino (mais específico). Esta perspectiva topólogica permite uma visão macro e micro dos padrões, tornando possível a análise de fatores de variabilidade que podem explicar ou afetar as repetições dos padrões e, portanto, a previsibilidade.

Na seção a seguir, cada um desses padrões será descrito em detalhe, a apresentação seguirá a estrutura definida pelo grafo, de modo a preservar associações e facilitar a compreensão da dinâmica entre os padrões.

4.1 Padrões descobertos

01 **MULTAS POR HORA E ANO**
Elementos Cada elemento é o somatório das multas agrupados por hora de um certo ano. Os anos considerados foram de 2020 até 2023, o que forma quatro elementos. As horas analisadas foram entre 6h e 21h. A cidade considerada foi Recife.
Regras Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto mínimo local. Às 9h, possui um ponto máximo local. Às 13h, possui um ponto mínimo local. Às 15h, possui o ponto de máximo global. A partir das 19h, possui o ponto de mínimo global.
Descrição O Gráfico 1, apresenta a quantidade de multas agrupadas por hora, que demonstra padrões temporais consistentes nas infrações de trânsito, entre os quatro anos considerados. Ao longo do dia os somatórios variam significativamente. A análise do gráfico identifica dois picos: um às 9h e outro pronunciado às 15h, separados por quedas abruptas às 12h e 18h.

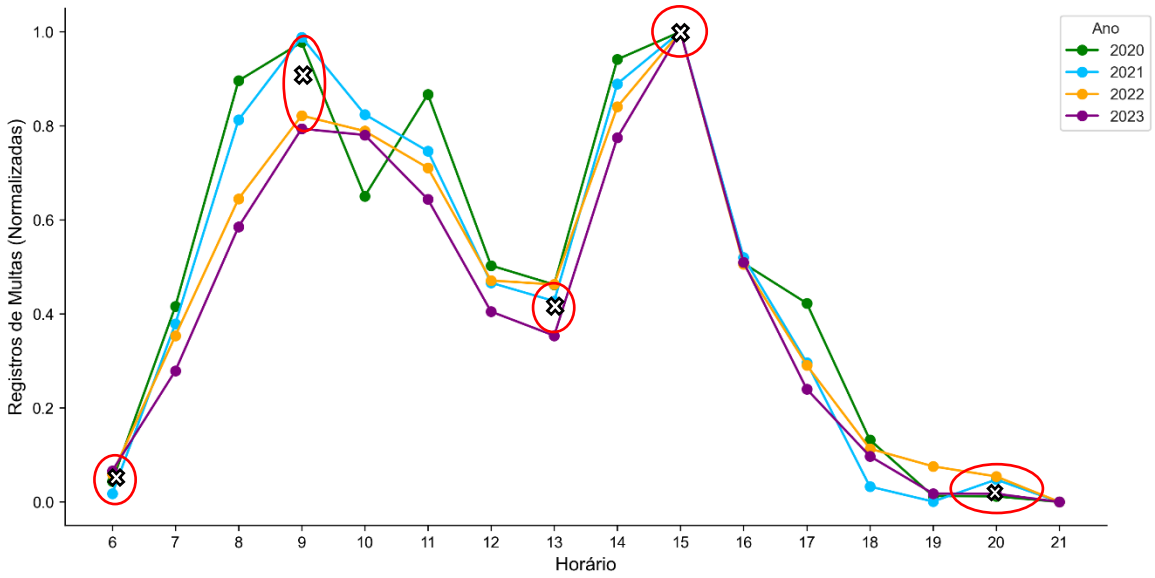
Gráfico 1 - Distribuição das multas por hora e ano



Fonte: Elaborado pelo autor.

A normalização dos dados Gráfico 2 confirma a estabilidade desses padrões entre os anos, indicando que as variações absolutas do Gráfico 1 não alteram a proporção relativa dos picos. Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra "X".

Gráfico 2 – Distribuição normalizada das multas por hora e ano



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética

O primeiro ponto de mínimo as 6h sugere que os motoristas estão em casa preparando-se para sair para seus compromissos, como trabalho e estudos. O segundo ponto de máximo local as 9h sugere um aumento de veículos nas ruas que consequentemente aumenta a probabilidade de cometimento de infrações. Os motoristas estão indo ao trabalho para iniciar o horário comercial as 9h, outros já estão trabalhando, visitando clientes, entregando produtos. O terceiro ponto de mínimo local as 13h sugere menos carros nas ruas devido ao horário de almoço. Em relação ao quarto ponto de máximo global as 15h, sugere o período de grande quantidade de veículos nas ruas devido ao retorno gradual do almoço e pico de movimento comercial. O quinto ponto da regra de mínimo global a partir das 19h sugere que os motoristas voltaram dos seus compromissos para casa.

A hipótese corrobora a ideia de que tanto o ambiente influencia o comportamento quanto o comportamento influencia o ambiente (Günther e Rozestraten, 2005). Além disso, Mohan (2002) e Zlapoter (1991) evidenciam que o aumento no fluxo de veículos afeta diretamente o ambiente do trânsito, contribuindo para o surgimento de situações de risco e elevação dos índices de infrações.

Associações

Este padrão relaciona-se com MULTAS POR HORA, MESES E ANO (02), MULTAS POR HORA, DIAS ÚTEIS DA SEMANA E ANO (03), MULTAS POR HORA, DIA DE SÁBADO E ANO (04), MULTAS POR HORA, DIA DE DOMINGO E ANO (05), MULTAS POR MÊS (06) e MULTAS DE VELOCIDADE POR HORA E ANO (07) para aprofundar o conhecimento sobre as multas de modo a descobrir em quais granularidades temporais este padrão se mantém, em quais não se mantém e quais são os fatores que afetam este padrão, tais como dia da semana, mês, pandemia, tipo de infração, política.

02
MULTAS POR HORA, MESES E ANO

Elementos

Cada elemento é o somatório das multas agrupados por hora e mês de um certo ano. Os anos considerados foram de 2020 até 2023 e todos os 12 meses de cada com exceção de abril e maio de 2020. O que forma 46 elementos. As horas consideradas foram entre 6h e 21h. A cidade foi recife.

Regras

Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto mínimo local. Às 9h, possui um ponto máximo local. Às 13h, possui um ponto mínimo local. Às 15h, possui o ponto de máximo local. A partir das 19h, possui o ponto de mínimo global.

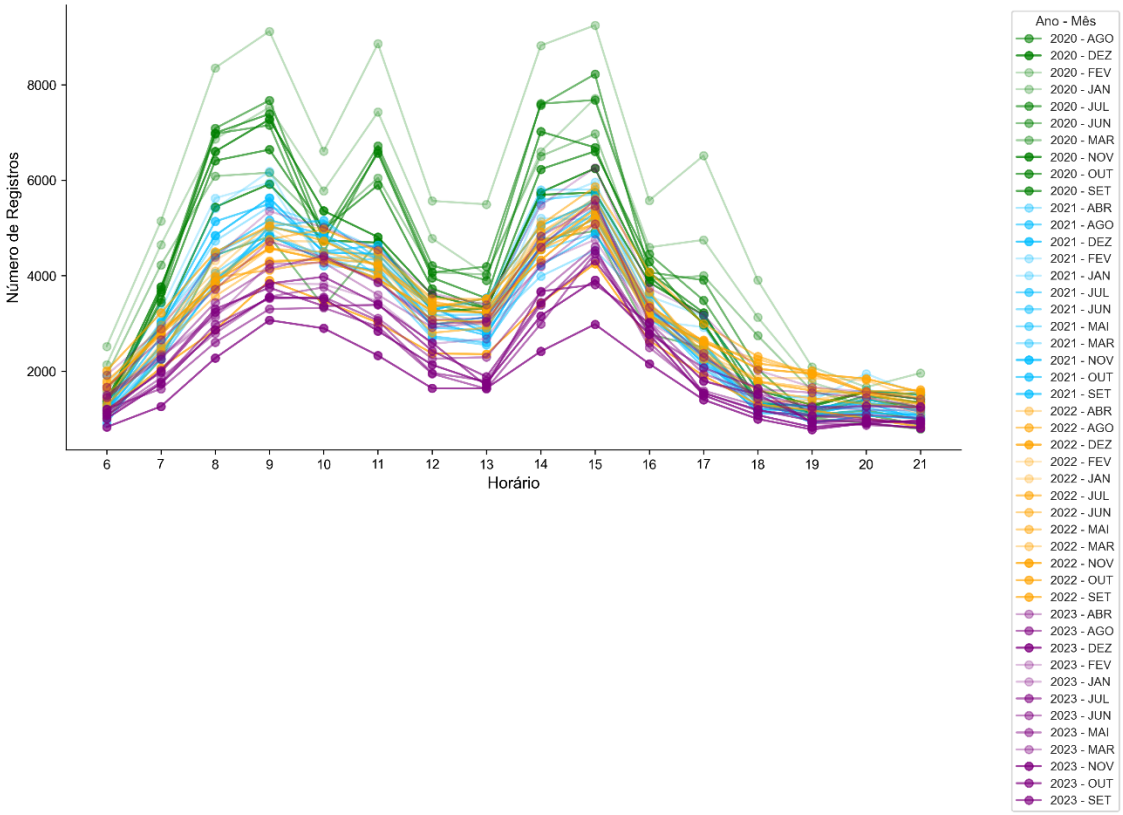
Descrição

O Gráfico 3 apresenta a distribuição absoluta do número de multas por hora, segmentada por mês e ano, com cada ano representado por uma cor específica. Com exceção dos meses de abril e maio de 2020, que foram impactados por causa da pandemia.

Em comparação ao padrão MULTAS HORA E ANO (01), as curvas dos anos são equivalentes e pode-se concluir que o comportamento das curvas se mantém.

No entanto, ao aumentar a especificidade do padrão MULTAS HORA E ANO (01) do ano para mês permitiu capturar maiores variações. Dois meses não seguem a regra do padrão MULTAS HORA ANO (01) devido a pandemia e vários meses possuem ponto de máximo local às 9h e não as 15h.

Gráfico 3 - Distribuição das multas por hora e mês

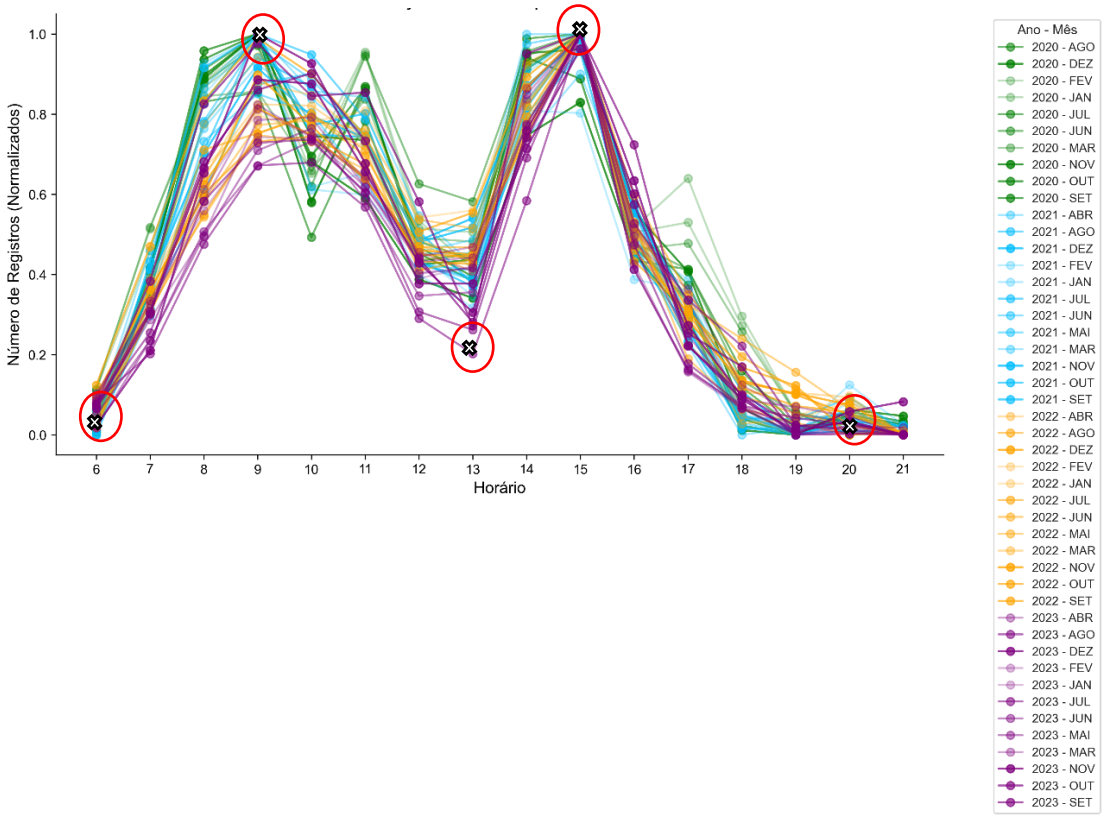


Fonte: Elaborado pelo autor, 2025.

Já o Gráfico 4 apresenta os dados em normalizados. Apesar das diferenças nos volumes absolutos de multas, o padrão temporal das infrações é consistente ao longo dos meses em relação aos anos. Os horários às 9h e as 15h é novamente identificado como um horário crítico para a ocorrência de multas. A normalização também destaca que os horários críticos para as os picos de infrações permanecem estáveis ao longo dos meses.

Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra "X".

Gráfico 4 - Distribuição normalizada das multas por hora e mês



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética

A análise da distribuição horária das multas gerais, considerando diferentes meses e anos, revela um padrão temporal estruturalmente consistente. As hipóteses levantadas do padrão MULTAS HORA ANO (01) podem ser consideradas também para o comportamento do padrão analisado.

Associações

Este padrão associa-se com MULTAS HORA E ANO (01) por adotar elementos mais específicos.

03
MULTAS POR HORA, DIAS ÚTEIS DA SEMANA E ANO

Elementos

Cada elemento é o somatório das multas agrupados por hora e dia da semana de um certo ano. Os anos considerados foram de 2020 até 2023. As horas consideradas foram entre 6h e 21h. Os 5 dias úteis de segunda à sexta, o que formam 20 elementos. A cidade foi recife.

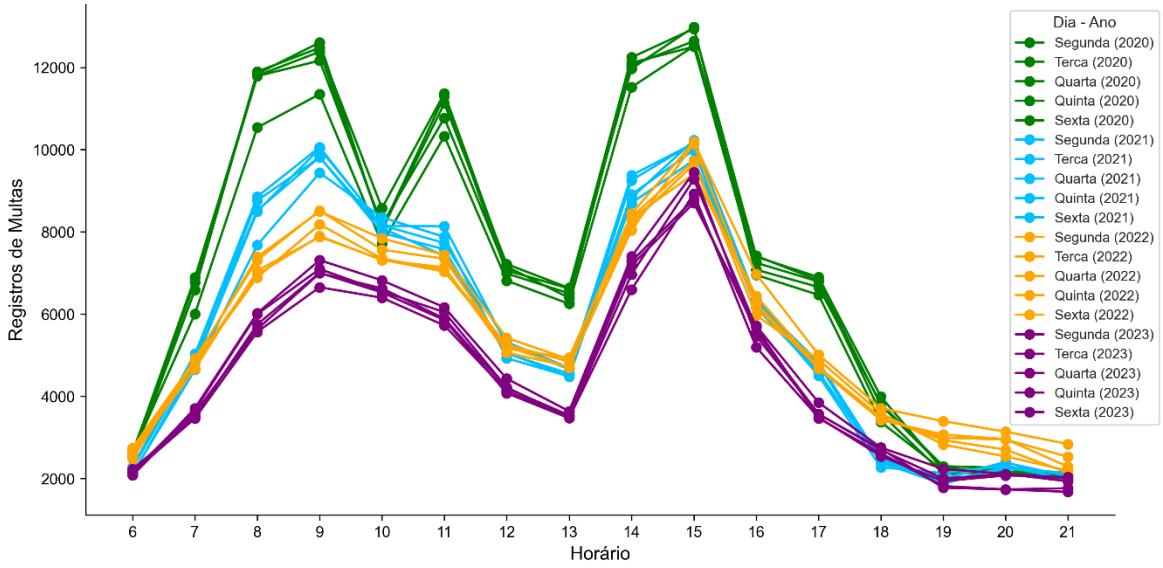
Regras

Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto mínimo local. Às 9h, possui um ponto máximo local. Às 13h, possui um ponto mínimo local. Às 15h, possui o ponto de máximo global. A partir das 19h, possui o ponto de mínimo local.

Descrição

O Gráfico 5 apresenta os registros absolutos de multas por hora em cada dia útil da semana. Observamos que, de segunda a sexta-feira, os picos de infrações ocorrem em horários específicos, um pico às 9h e outro a tarde às 15h, alinhados aos comportamentos observados no padrão MULTAS HORA E ANO (01).

Gráfico 5 - Distribuição das multas por hora, dias úteis da semana e ano

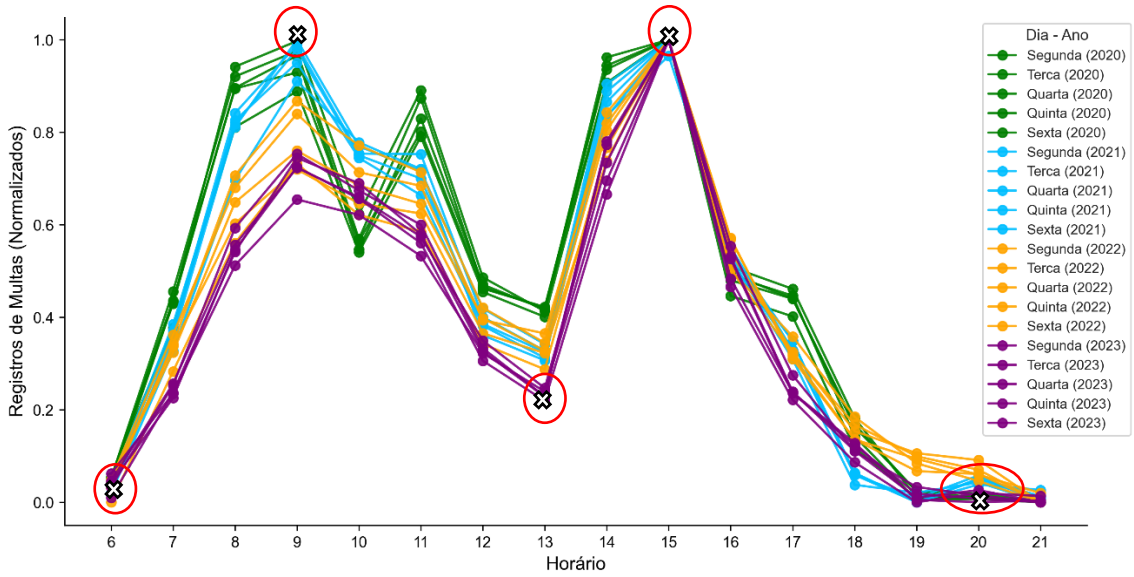


Fonte: Elaborado pelo autor, 2025.

A normalização dos dados Gráfico 5 confirma a estabilidade desses padrões entre os anos, indicando que as variações absolutas do Gráfico 6 não alteram a proporção relativa dos picos. As 9h e às 15h é novamente identificado como um horário crítico para a ocorrência de multas durante os dias úteis.

Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra "X".

Gráfico 6 - Distribuição normalizada das multas por hora, dias úteis da semana e ano



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética

A análise da distribuição horária das multas gerais, considerando os dias úteis da semana e anos, revela um padrão temporal estruturalmente consistente. As hipóteses levantadas do padrão MULTAS HORA ANO (01) podem ser consideradas também para o comportamento do padrão analisado. Dessa forma, sugere-se que fatores atrelados a rotina e compromissos dos condutores influenciam o comportamento dos registros para os dias da semana.

Associações

Este padrão associa-se com MULTAS HORA E ANO (01) por adotar elementos mais específicos.

04
MULTAS POR HORA, SÁBADO E ANO

Elementos

Cada elemento é o somatório das multas agrupados por hora de um certo ano ocorridos no sábado. As horas consideradas foram entre 6h e 21h. Os anos considerados foram de 2020 até 2023, o que forma 4 elementos.

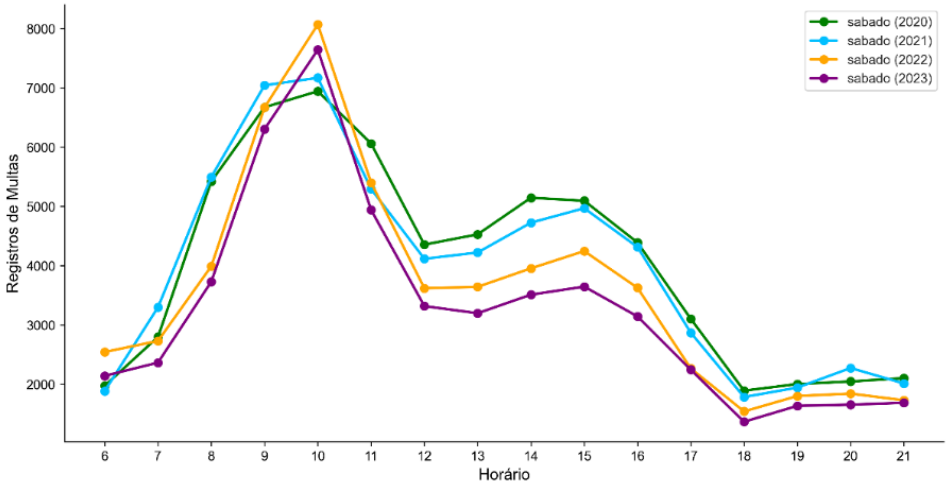
Regras

Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto mínimo local. Às 10h, possui um ponto máximo local. Entre 12h e 13h, possui um ponto mínimo local. Às 15h, possui o ponto de máximo local. A partir das 18h, possui o ponto de mínimo global.

Descrição

O Gráfico 7 exibe os registros absolutos de multas por hora aos sábados. Observamos que os anos apresentam um pico bem definido às 10h apesar da divergência entre os valores absolutos entre os anos analisados. A queda nas infrações ocorre de forma gradual após as 15h, independentemente do ano analisado.

Gráfico 7 - Distribuição das multas ocorridas no sábado por hora e ano

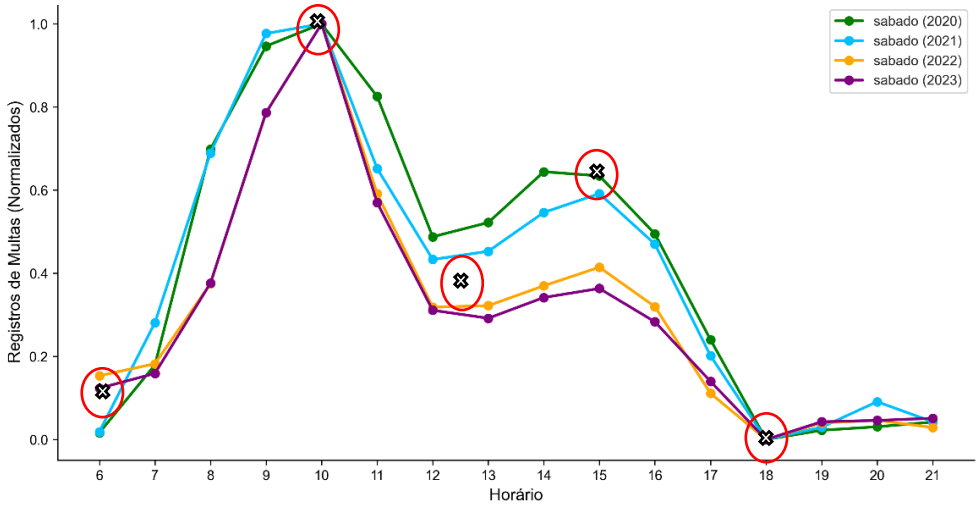


Fonte: Elaborado pelo autor, 2025.

O Gráfico 8 apresenta a versão normalizada dos dados, ajustando os valores absolutos para refletir a proporção relativa de infrações em cada horário. A curva normalizada também indica que, os horários críticos para ocorrência das multas se mantêm relativamente estáveis ao longo dos anos.

Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra "X".

Gráfico 8 - Distribuição normalizada das multas ocorridas no sábado por hora e ano



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética
O ponto de máximo local às 6h sugere menor movimento de carro nas ruas pois a maioria dos motoristas ainda estão em casa. O ponto máximo global às 10h sugere um comportamento similar ao padrão MULTAS POR HORA, DIAS ÚTEIS DA SEMANA E ANO (03) pelo movimento comercial no sábado que tipicamente ocorre até às 12h. Com queda na hora do almoço. O ponto de máximo local às 15h significativamente menor que às 10h sugere menor movimentação comercial. Finalizando, às 18h com os motoristas retornando para casa.

Associações
Este padrão associa-se com MULTAS HORA E ANO (01) por adotar elementos mais específicos.

05
MULTAS POR HORA, DOMINGO E ANO

Elementos
Cada elemento é o somatório das multas agrupados por hora de um certo ano ocorridos no domingo. As horas consideradas foram entre 6h e 21h. Os anos considerados foram de 2020 até 2023, o que forma 4 elementos.

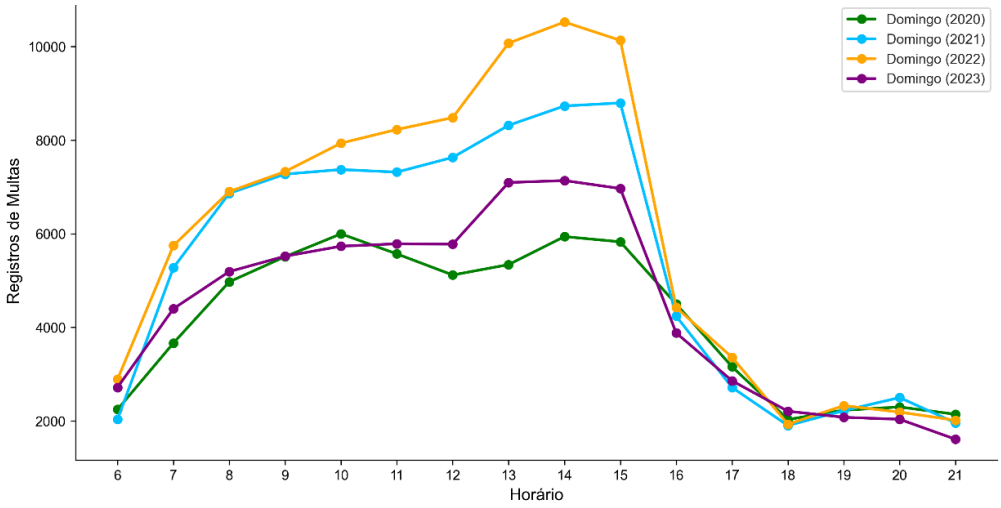
Regras
Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto mínimo local. Entre 14h e 15h, possui o ponto de máximo local. A partir das 18h, possui o ponto de mínimo global.

Descrição
O Gráfico 9 exibe os registros absolutos de multas por hora ocorridas aos domingos. Observa-se que a partir das 6h os anos apresentam um aumento gradual e um pico bem definido com início às 13h, apesar da divergência entre os valores absolutos entre os anos analisados.

A queda nas infrações ocorre de forma acentuada após as 15h, independentemente do ano analisado.

Esse padrão contrasta com o padrão MULTAS POR HORA, DIAS ÚTEIS DA SEMANA E ANO (03), nos quais os picos de infrações ocorrem durante períodos de deslocamento para o trabalho, e com o padrão MULTAS POR HORA, SÁBADO E ANO (04), que apresentam um comportamento de infração mais concentrado no período da manhã.

Gráfico 9 - Distribuição das multas ocorridas no domingo por hora e ano

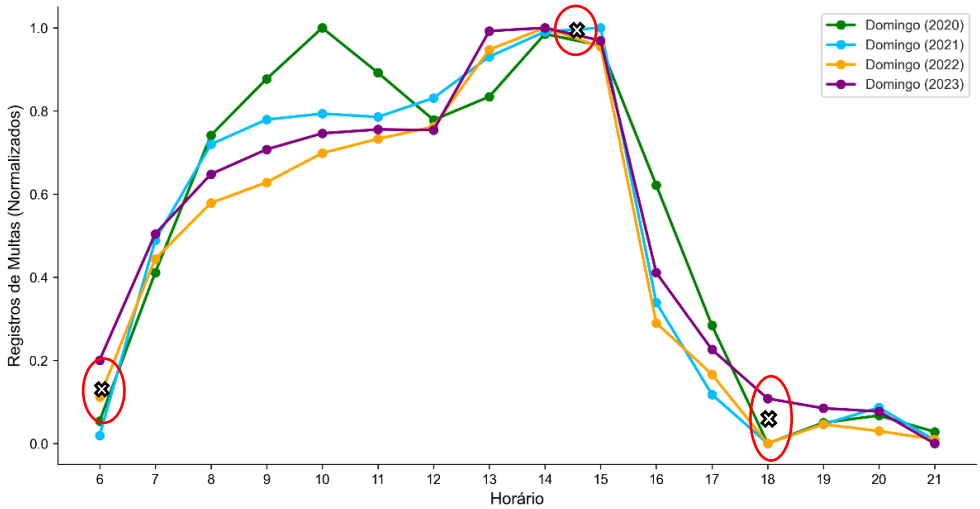


Fonte: Elaborado pelo autor, 2025.

A normalização dos dados Gráfico 10 ajusta os valores absolutos para refletir a proporção relativa de infrações em cada horário. A normalização indica que as variações absolutas do Gráfico 9 altera a proporção relativa do pico para às 10h no ano de 2020, porém o pico entre às 14h e 15h se mantém estável para os anos mais recentes.

Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra "X".

Gráfico 10 - Distribuição normalizada das multas ocorridas no domingo por hora e ano



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética
A análise monológica da distribuição das infrações de trânsito aos domingos revela um padrão diferenciado em relação aos dias úteis e aos sábados. O aumento gradual das multas a partir das 6h ao longo da manhã, culminando em um pico às 14h e 15, sugere que a dinâmica de deslocamento típicos desse dia pode estar associada a atividades de lazer e turismo urbano. A partir das 18h reduzindo-se progressivamente o número de carros conforme essas atividades se encerram.

Outro aspecto relevante é o volume absoluto de infrações aos domingos do ano de 2020, que que possui pico de registros superior às 10h. Sugere que esse comportamento pode ser devido a pandemia. A partir dos dados analisados, observa-se que, embora o volume total de infrações tenha diminuído em anos mais recentes, a estrutura temporal das infrações se mantém constante.

Associações
Este padrão associa-se com MULTAS HORA E ANO (01) por adotar elementos mais específicos.

06
MULTAS POR MÊS

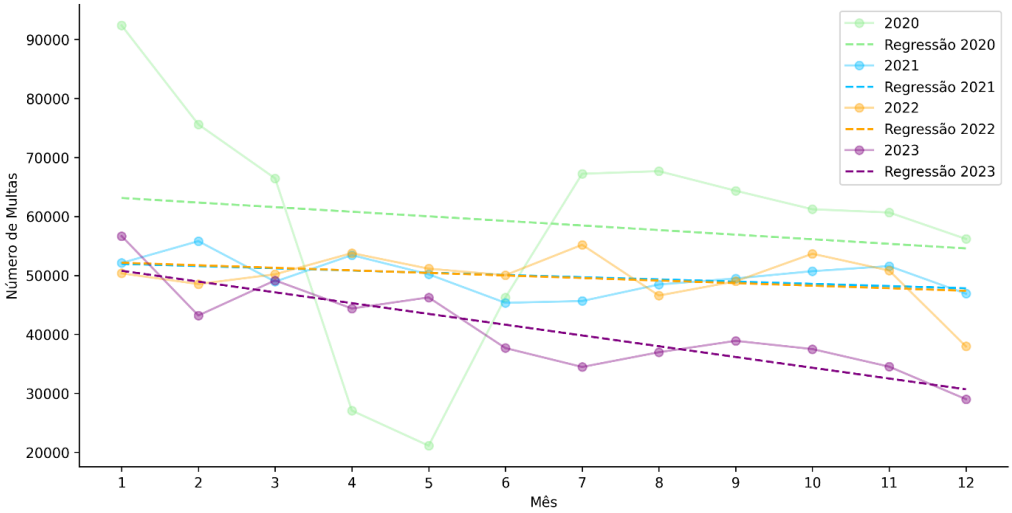
Elementos
Cada elemento é o somatório das multas agrupados por meses de um certo ano. Os anos considerados foram de 2020 até 2023, o que forma 12 elementos.

Regras
A regressão linear dos elementos possui desvio angular negativo.

Descrição
O Gráfico 11 apresenta a regressão linear para os anos de 2020 a 2023, representada por linhas tracejadas, com o eixo X indicando os meses do ano e o eixo Y representando o número de multas registradas. Cada linha tracejada corresponde à tendência linear dos dados de infrações em cada ano, enquanto os pontos conectados por linhas sólidas mostram os valores absolutos mensais.

Este padão considera infrações relacionadas categorizando sua incidência ao longo dos meses. Observa-se que as regressões lineares apresentam desvios angulares negativos, evidenciando uma tendência geral de redução no número de multas ao longo dos meses.

Gráfico 11 - Regressão linear simples das multas por mês



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética

A tendência identificada revela uma diminuição progressiva e significativa na incidência de infrações ao longo do ano, com declínio mais evidente no último trimestre, especialmente no mês de dezembro. Esse comportamento pode estar associado a fatores sazonais, como a redução do tráfego devido às férias escolares e recessos.

O ano de 2020 apresenta uma queda abrupta durante os meses de abril e maio, provavelmente devido às restrições de mobilidade impostas pela pandemia da COVID-19, que reduziram significativamente a circulação de veículos e, consequentemente, as infrações. Já em 2021, 2022 apresentam comportamento mais próximo, sugerindo uma normalização do fluxo veicular e das condições de trânsito nesses períodos.

O ano de 2023 apresenta número reduzido de infrações, devido ao desligamento de radares. De acordo com CTTU https://cttu.recife.pe.gov.br/, a licitação dos aparelhos venceu em junho de 2023 culminando em 25 radares desativados na cidade.

A regressão linear evidencia um possível padrão de declínio das infrações ao longo do ano, mas que a hipótese pode ser também uma coincidência influenciado por fatores externos e políticos.

Associações

Este padrão relaciona-se com MULTAS POR HORA, MESES E ANO (02) por adotar elementos mais genéricos.

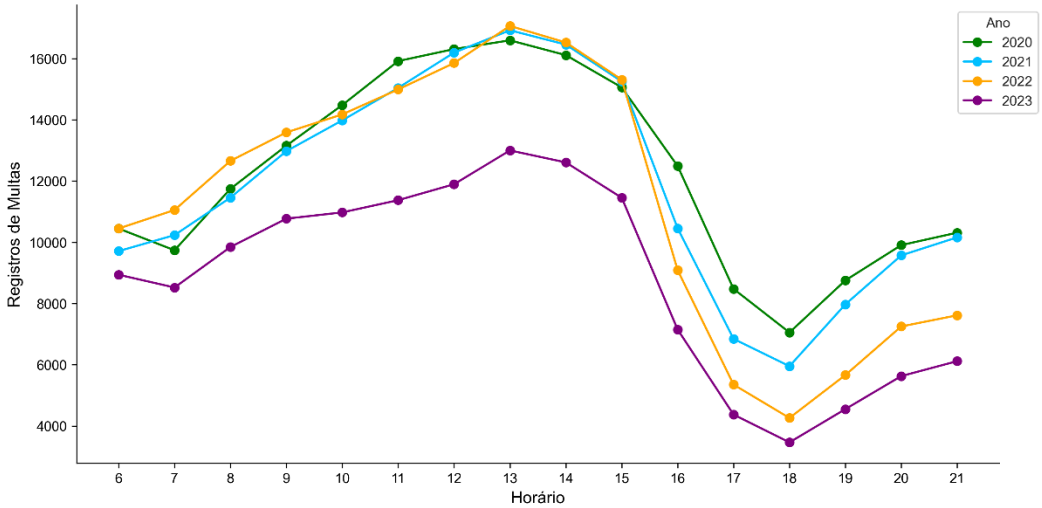
07
MULTAS DE VELOCIDADE POR HORA E ANO

Elementos
Cada elemento é o somatório das multas de infrações de velocidade acima de 20% agrupados por hora de um certo ano. As horas consideradas foram entre 6h e 21h. Os anos considerados foram de 2020 até 2023, o que forma 4 elementos.

Regras
Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto mínimo local. Às 13h, possui o ponto de máximo global. Às 18h, possui o ponto de mínimo global. Às 21h, um ponto máximo local.

Descrição
O Gráfico 12 representa os valores absolutos, observa-se um aumento gradual no número de infrações a partir das 6h da manhã, atingindo seu pico às 13h. Após esse horário, há uma queda acentuada no volume de infrações, com os menores registros concentrados entre as 18h e o final da noite. Apesar das diferenças nos valores absolutos entre os meses e anos analisados, o padrão temporal de comportamento se mantém estável.

Gráfico 12 - Distribuição das infrações de velocidade acima de 20% por hora e ano

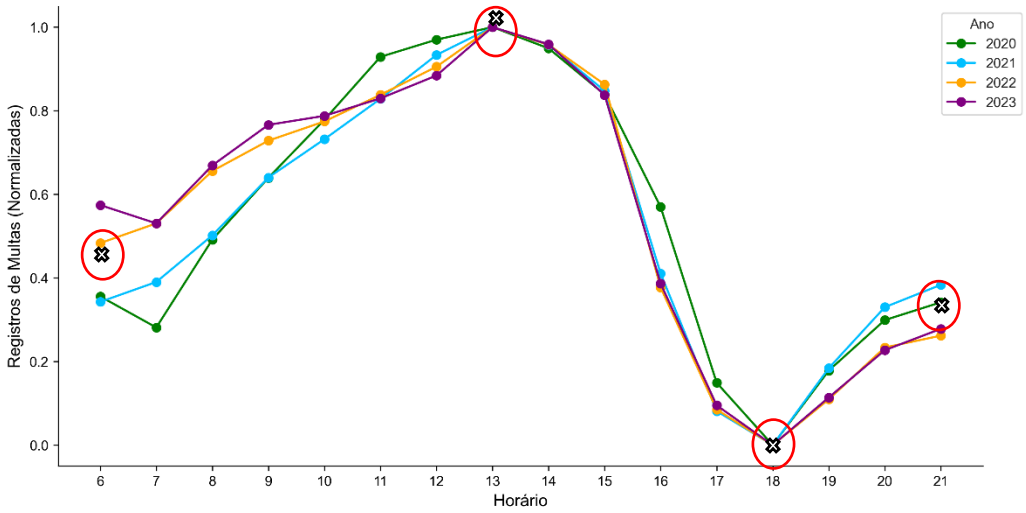


Fonte: Elaborado pelo autor, 2025.

A normalização dos dados Gráfico 13 confirma a estabilidade desses padrões entre os anos, indicando que as variações absolutas do Gráfico 12 não alteram a proporção relativa dos picos.

Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra "X".

Gráfico 13 - Distribuição normalizada das infrações de velocidade acima de 20% por hora e ano



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética.

Esse padrão sugere uma relação direta entre fluxo veicular, comportamento do condutor e fiscalização Mohan (2002) e Zlapoter (1991). Durante a manhã, o crescimento gradual das infrações pode estar relacionado ao desbloqueio gradativo do trânsito depois do horário de pico, período em que os condutores estão se deslocando para compromissos, trabalhos e estudos. Dessa forma, permitindo que condutores acelerem além do limite permitido. O pico observado às 13h pode ser justificado pela menor densidade de veículos nesse período, tendo em vista o horário de almoço, facilitando comportamentos de risco, como o excesso de velocidade.

Diferente do padrão MULTAS POR HORA E ANO (01), que apresentam picos mais distribuídos ao longo do dia, as infrações de velocidade seguem um crescimento gradual ao longo da manhã até atingirem o pico no período de almoço. A queda expressiva às 18h sugere que o trânsito mais intenso nos horários de maior movimentação comercial que impede fisicamente que os motoristas excedam a velocidade permitida (Feitosa, 2010).

Esse comportamento pode ser reforçado pela maior concentração de veículos, pelo receio de colisões e pelo aumento da fiscalização em determinados pontos da cidade (Gunther, 2001).

O leve aumento das infrações após as 19h pode indicar que após horários de pico, motoristas percebem as vias menos congestionadas e, consequentemente, adotam comportamentos mais arriscados, como acelerações bruscas e velocidades acima do permitido (Rozestraten,1988) e (Gunther, 2001).

Associações

Este padrão relaciona-se com MULTAS DE VELOCIDADE POR HORA, MESES E ANO (08), MULTAS DE VELOCIDADE POR HORA, DIAS ÚTEIS DA SEMANA E ANO (09), MULTAS DE VELOCIDADE POR HORA, FINAL DE SEMANA E ANO (10) e MULTAS DE VELOCIDADE POR INTENSIDADE DE CHUVA E ANO (11) para analisar com maior profundidade os padrões das multas de infrações de velocidade, analisando suas especificidades e identificando em quais escalas

temporais como dias da semana, meses ou períodos específicos esses padrões se mantêm ou se alteram, além de investigar os fatores que influenciam essas variações, como a pandemia e eventos sazonais.

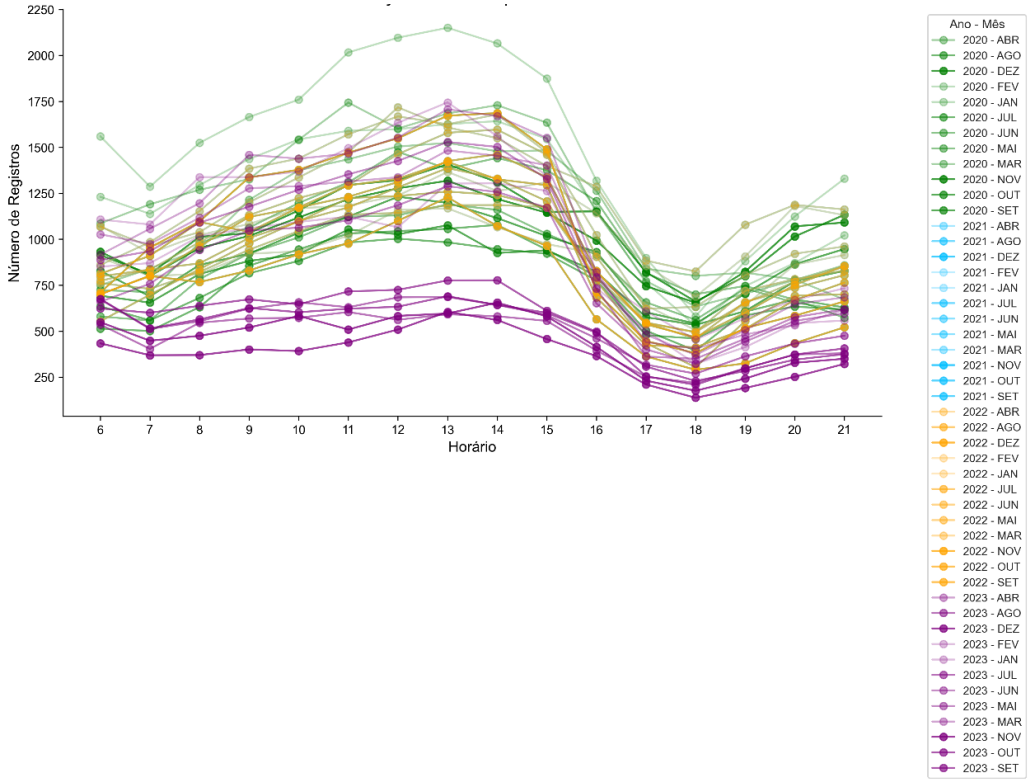
08
MULTAS DE VELOCIDADE POR HORA, MESES E ANO

Elementos
Cada elemento é o somatório das multas de infrações de velocidade acima de 20% agrupados por hora e meses de um certo ano. As horas consideradas foram entre 6h e 21h. Os anos considerados foram de 2020 até 2023, o que forma 48 elementos.

Regras
Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto máximo local. Às 7h, possui um ponto mínimo local. Às 12h, possui o ponto de máximo global. Às 18h, possui o ponto de mínimo global. Às 21h, possui um ponto máximo local.

Descrição
O gráfico 14, exibe os valores absolutos das infrações de velocidade, percebe-se um padrão consistente entre os meses ao longo dos anos. Ao aumentar a especificidade do padrão MULTAS DE VELOCIDADE POR HORA E ANO (07) do ano para mês permitiram observar maiores variações nas curvas. Alguns meses possuem ponto de máximo local entre 12h e 14h.

Gráfico 14 - Distribuição de infrações de velocidade acima de 20% por hora e mês

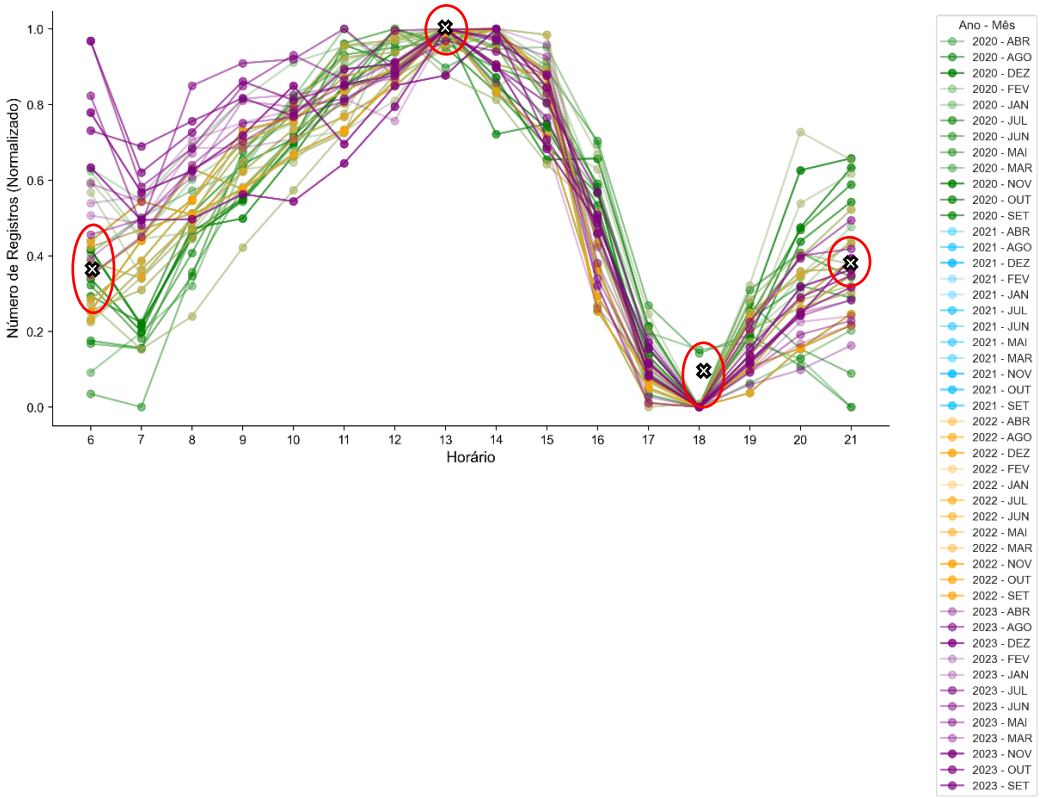


Fonte: Elaborado pelo autor.

O Gráfico 15 normalizado expõe em termos relativos a semelhança de comportamento entre os meses durante os anos.

Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra “X”.

Gráfico 15 - Distribuição normalizada de infrações de velocidade acima de 20% por hora e mês



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética

A análise da distribuição horária das infrações de velocidade acima de 20%, considerando diferentes meses e anos, revela um padrão temporal estruturalmente consistente. As hipóteses levantadas para o padrão MULTAS DE VELOCIDADE POR HORA E ANO (07) podem ser consideradas também para o comportamento do padrão analisado.

Tal hipótese evidencia que o comportamento infracional analisado, de modo geral não está sujeito a variações sazonais.

Associações

Este padrão associa-se indiretamente com MULTAS HORA E ANO (01) e diretamente com MULTAS DE VELOCIDADE POR HORA E ANO (07) por adotar elementos mais específicos.

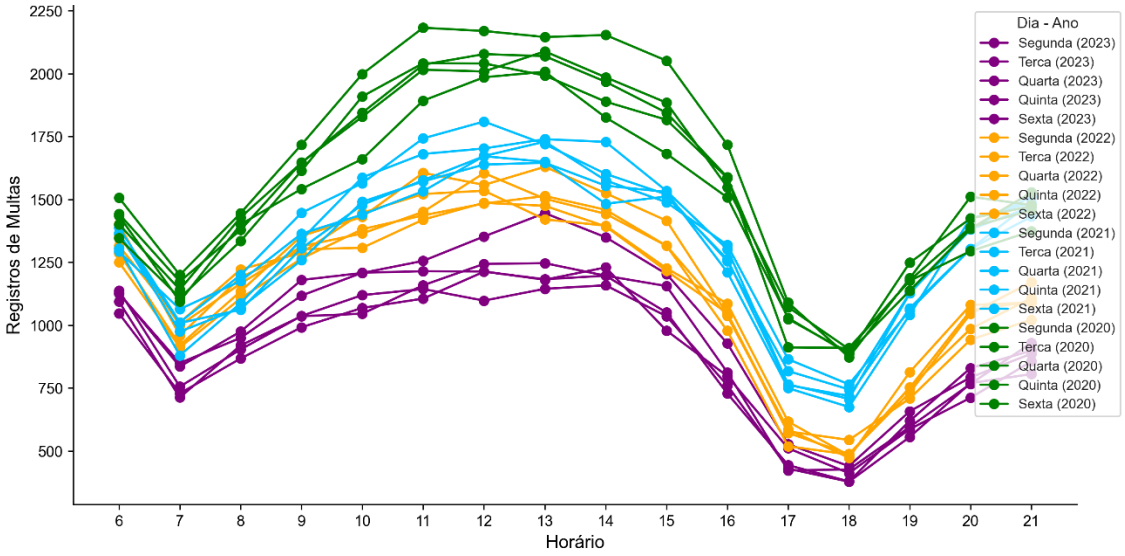
09
MULTAS DE VELOCIDADE POR HORA, DIAS ÚTEIS DA SEMANA E ANO

Elementos
Cada elemento é o somatório das multas de infrações de velocidade acima de 20% agrupados em valores absolutos por hora, dias úteis da semana e ano. As horas consideradas foram entre 6h e 21h. Os anos agrupados por dias úteis foram de 2020 até 2023, o que forma 20 elementos.

Regras
Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto máximo local. Às 7h, possui um ponto mínimo local. Entre às 11h às 14h, possui o ponto de máximo local. Às 18h, possui o ponto de mínimo global. Após 18h, possui um ponto máximo local.

Descrição
O Gráfico 16 apresenta os registros absolutos de multas por hora em cada dia útil da semana. Observamos que, de segunda a sexta-feira, o comportamento das curvas se mantém estáveis do padrão MULTAS DE VELOCIDADE POR HORA E ANO (07). Mas a partir do gráfico podemos observar que após aumentar a especificidade dos elementos o intervalo de picos de registros é maior, abrangendo os horários das 11h às 14h.

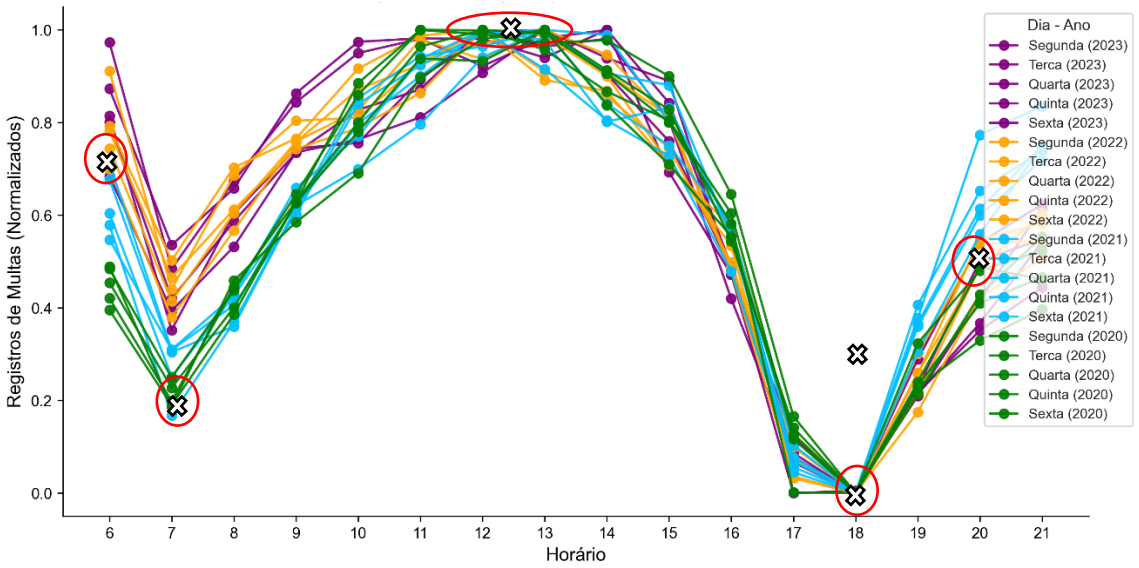
Gráfico 16 - Distribuição de infrações de velocidade acima de 20% por hora, dia útil e ano



Fonte: Elaborado pelo autor, 2025.

A normalização dos dados Gráfico 17 confirma a estabilidade desses padrões entre os anos, indicando que as variações absolutas dos dias úteis do Gráfico 16 não alteram a proporção relativa dos picos.
Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra "X".

Gráfico 17 - Distribuição normalizada de infrações de velocidade acima de 20% por hora, dia útil e ano



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética

A análise da distribuição horária das infrações de velocidade acima de 20%, considerando diferentes meses e anos, revela um padrão temporal estruturalmente consistente. As hipóteses levantadas para o padrão MULTAS DE VELOCIDADE POR HORA E ANO (07) podem ser consideradas também para o comportamento do padrão analisado.

A consistência dos padrões ao longo dos dias úteis e anos reforça a hipótese de que há uma estrutura previsível nos comportamentos infracionais.

Associações

Este padrão associa-se indiretamente com MULTAS HORA E ANO (01) e diretamente com MULTAS DE VELOCIDADE POR HORA E ANO (07) por adotar elementos mais específicos.

10
MULTAS DE VELOCIDADE POR HORA, FINAL DE SEMANA E ANO

Elementos

Cada elemento é o somatório das multas de infrações de velocidade acima de 20% agrupados em valores absolutos por hora, dias final de semana e ano. As horas consideradas foram entre 6h e 21h. Os anos agrupados por final de semana (sábado e domingo) foram de 2020 até 2023, o que forma 8 elementos.

Regras

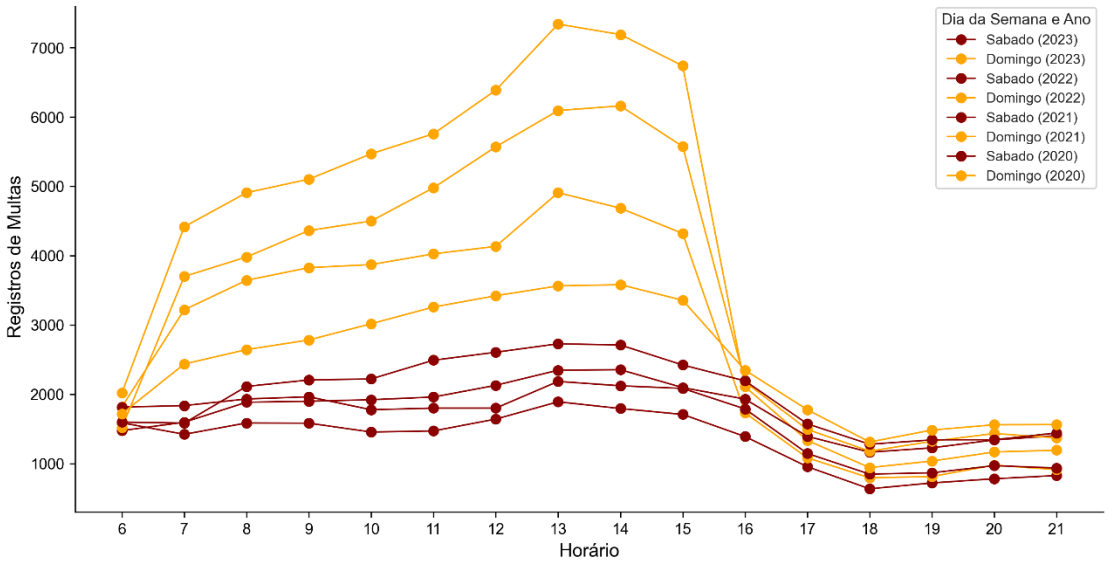
Cada elemento possui os seguintes pontos mínimos e máximos locais. Às 6h, possui um ponto mínimo local. Às 13h, possui um ponto máximo global. A partir das 18h, possui o ponto de mínimo global.

Descrição

O Gráfico 18, exibe os valores absolutos das infrações, observa-se que os dias de final de semana, possuem divergências em termos de volumes de registros de infrações de velocidade, evidenciado pelo destaque da curva em laranja (sábado) em relação ao e vermelho-escuro (domingo).

O domingo parece apresentar um padrão distinto em relação ao sábado, com um aumento significativo no número de infrações durante a manhã e o início da tarde (13h), atingindo um pico maior em termos absolutos.

Gráfico 18 - Distribuição de infrações de velocidade acima de 20% por hora, final de semana e ano

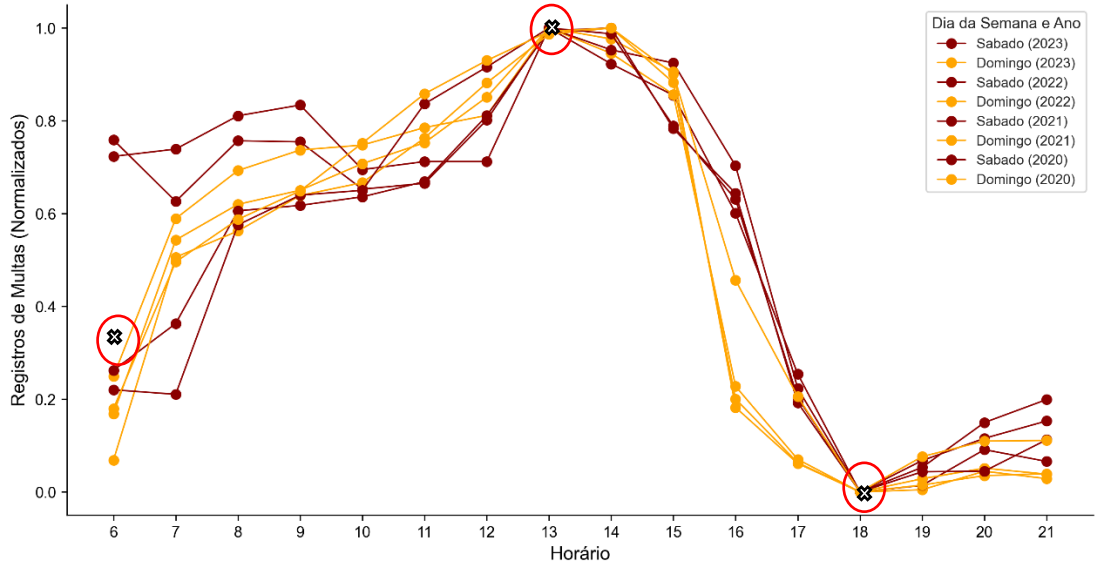


Fonte: Elaborado pelo autor, 2025.

No entanto, ao normalizar esses dados no Gráfico 19, observamos que em proporções relativas os dias possuem curvas semelhantes. Assim, parece que sábado e domingo seguem um provável padrão semelhante de crescimento e declínio das infrações, com uma possível tendência de aumento ao longo da manhã até às 13h.

Os pontos mínimos e máximos locais da regra estão representados no Gráfico 2 pela letra "X".

Gráfico 19 - Distribuição de infrações de velocidade acima de 20% por hora, final de semana e ano



Fonte: Elaborado pelo autor, 2025.

Explicação Hipotética

Os padrões de infrações de velocidade nos finais de semana mostram que tanto sábado quanto domingo seguem um comportamento semelhante. A partir das 6h ocorre aumento das infrações com pico por volta das 13h, seguido de queda acentuada. Às 18h sugere que os condutores voltaram de suas atividades para casa.

A variação nos registros durante a manhã aos sábados, sugere uma possível influência do horário comercial.

O domingo se destaca com maior volume absoluto, que possivelmente pode estar ligado a atividades de lazer e uma possível percepção reduzida de fiscalização. De forma geral, o comportamento descrito tende a confirmar os padrões já evidenciados anteriormente nas análises.

Associações

Este padrão associa-se indiretamente com MULTAS HORA E ANO (01) e diretamente com MULTAS DE VELOCIDADE POR HORA E ANO (07) por adotar elementos mais específicos.

5 CONSIDERAÇÕES FINAIS

A principal contribuição deste trabalho é o descoberta e análise de padrões sobre infrações de trânsito do Recife, aplicando técnicas de mineração de dados em bases de dados disponibilizadas pelo Portal de Dados Abertos do Recife.

Os resultados demonstraram que há padrões temporais consistentes das infrações de trânsito ao longo do dia, independentemente de variáveis sazonais. Tais achados confirmam que a análise de dados públicos permite não apenas compreender com maior profundidade as dinâmicas da mobilidade urbana, como também fundamentar o planejamento estratégico de políticas públicas voltadas à segurança viária, mobilidade e infraestrutura urbana.

A principal contribuição deste trabalho reside na demonstração prática de como a mineração de dados públicos pode servir como ferramenta para a tomada de decisões mais embasadas, transparentes e eficientes na gestão urbana. Além de reforçar a importância da transparência e da disponibilidade de dados governamentais, o estudo aponta para o potencial do processo de KDD como aliada na formulação de políticas públicas baseadas em evidências.

Entre as principais limitações enfrentadas neste estudo, destaca-se a limitação temporal e espacial dos dados disponíveis. Em alguns casos, houve dificuldade na obtenção de dados infracionais mais recentes como os de 2024, bem como restrições na padronização e na qualidade dos dados abertos utilizados relacionados a local de cometimento das multas, para um possível georreferenciamento. Além disso, o escopo metodológico centrou-se em variáveis específicas de infrações, o que não contemplou outros fatores relevantes à dinâmica urbana.

Para estudos futuros, recomenda-se a ampliação do escopo analítico, incluindo outras fontes de dados públicas ou privadas que possam complementar a análise, como informações sobre acidentes, transporte coletivo, e sensores urbanos e dados de infraestrutura viária. Além disso, a aplicação de modelos preditivos baseados em algoritmos de aprendizado de máquina representa uma frente promissora para a antecipação de comportamentos futuros. Tais modelos podem ser utilizados para prever padrões de infração, congestionamentos e impactos de eventos climáticos em diferentes cenários urbanos, contribuindo para o planejamento preventivo e para a gestão proativa da mobilidade urbana.

Por fim, recomenda-se um aprofundamento teórico e aplicado nas discussões relacionadas à gestão do conhecimento baseada em padrões, considerando que os padrões extraídos por meio de técnicas de mineração de dados não apenas descrevem comportamentos recorrentes, mas também podem ser formalizados, reaproveitados e comunicados como ativos de conhecimento organizacional.

REFERÊNCIAS BIBLIOGRÁFICAS

ALEXANDER, Christopher et al. ***A Pattern Language: Towns, Buildings, Construction***. New York: Oxford University Press, 1977.

AGGARWAL, Charu C. ***Data Mining: The Textbook*.** Cham: Springer, 2015.
BOURDIEU, Pierre. *O senso prático*. Petrópolis: Vozes, 1980.

AGRAWAL, Rakesh; IMIELINSKI, Tomasz; SWAMI, Arun. ***Mining association rules between sets of items in large databases***. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1993. p. 207-216. Disponível em: https://dl.acm.org/doi/10.1145/170036.170072. Acesso em: 20 set. 2024.

ALI, Peshawa Jamal Muhammad; FARAJ, Rezhna Hassan. ***Data normalization and standardization: a technical report***. Machine Learning Technical Reports, Koya University, v. 1, n. 1, p. 1–6, 2014. Disponível em: https://www.researchgate.net/publication/340579135_Data_Normalization_and_Standardization_A_Technical_Report. Acesso em: 30 mar. 2025.

AZAD, S. A., WASIMI, S., & Ali, A. B. M. S. (2018). ***Business Data Enrichment: Issues and Challenges***. 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE). doi:10.1109/apwconcse.2018.00024.

BETTENCOURT, Luís M. A.; WEST, Geoffrey B. ***A unified theory of urban living***. Nature, London, v. 467, p. 912–913, 2010. Disponível em: https://www.researchgate.net/publication/47510517_A_unified_theory_of_urban_living. Acesso em: 29 mar. 2025.

BARROS, Aidil J. da Silveira; LEHFELD, Neide Aparecida de souza. **Fundamentos de Metodologia científica**. 3. ed. São Paulo: Pearson Prentice Hall, 2007.

BARROS, A. J. P; LEHFELD, N. A. de S. **Fundamentos de metodologia: um guia para a iniciação científica**. 2. ed. ampliada. São Paulo: Makron Books, 2000.

BRASIL. Lei nº 9.503, de 23 de setembro de 1997. **Institui o Código de Trânsito Brasileiro**. Diário Oficial da União: seção 1, Brasília, DF, ano 135, n. 184-E, p. 1-34, 24 set. 1997. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9503.htm. Acesso em: 1 abr. 2025

BRASIL. Decreto n.º 7.724, de 16 de maio de 2012. **Regulamenta a Lei n.º 12.527, de 18 de novembro de 2011, que dispõe sobre o acesso a informações previsto no inciso XXXIII do caput do art. 5.º, no inciso II do § 3.º do art. 37 e no § 2.º do art. 216 da Constituição Federal**. Diário Oficial da União: seção 1, Brasília, DF, 17 maio 2012. Disponível em: https://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Decreto/D7724.htm. Acesso em: 15 set. 2024.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011**. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal**. *Diário Oficial da União: seção 1*, Brasília, DF, 18 nov. 2011. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 29 mar. 2025.

BRASIL. Decreto n.º 8.777, de 11 de maio de 2016. **Institui a Política de Dados Abertos do Poder Executivo Federal**. Diário Oficial da União: seção 1, Brasília, DF, 12 maio de 2016. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm. Acesso em: 15 set. 2024.

BRASIL. **Carta brasileira para cidades inteligentes. Brasília: Ministério do Desenvolvimento Regiona**l; Ministério da ciência, tecnologia e inovações (MCTI), 2022. 196 p. Disponível em: https://www.gov.br/cidades/pt-br/acesso-a-informacao/acoes-e-programas/desenvolvimento-urbano-e-metropolitano/projeto-andus/carta-brasileira-para-cidades-inteligentes. Acesso em: 01 abr. 2025.

COOPER, Donald R.; SCHINDLER, Pamela. **Métodos de Pesquisa em Administração**. 7. ed. Porto Alegre: Bookman, 2003.

CRONBACH, L. J.; MEEHL, P. E. ***Construct validity in psychological tests***. Psychological Bulletin, v. 52, p. 281–302, 1955. Disponível em: http://dx.doi.org/10.1037/h0040957. Acesso em: 30 mar. 2025.

CAPURRO, Rafael; HJØRLAND, Birger. ***The concept of information***. *Annual Review of Information Science and Technology*, v. 37, n. 1, p. 343–411, 2003.

EDELSTEIN, Herbert A. ***Introduction to data mining and knowledge discovery***. 2. ed. Two Crows Corporation, 1998.

ELIAS, Norbert. **O processo civilizador: uma história dos costumes**. 2. ed. Rio de Janeiro: Jorge Zahar, 1994.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. ***Knowledge discovery and data mining: towards a unifying framework***. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 1996.

FEITOSA, Zuleide Oliveira. **Competição por espaço em estacionamento público: invasão, reações e justificativas diante de vagas reservadas**. 2010. 65 f. Dissertação (Mestrado em Psicologia Social do Trabalho e das Organizações) – Universidade de Brasília, Brasília, 2010.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data Mining: um guia prático**. Elsevier, 2005. ISBN 978-8535218770.

GIDDENS, Anthony**.** ***The Constitution of Society: Outline of the Theory of Structuration***. Cambridge: Polity Press, 1984.

GEERTZ, Clifford. **A interpretação das culturas**. Rio de Janeiro: LTC, 1989. (Obra original publicada em 1973).

GARFINKEL, Harold. ***Studies in ethnomethodology***. Englewood Cliffs: Prentice-Hall, 1967.

GÜNTHER, H.; ROZESTRATEN. **Psicologia ambiental: algumas considerações sobre sua área de pesquisa e ensino** (Série: Textos de Psicologia Ambiental, n. 10). Brasília, DF: UnB, Laboratório de Psicologia Ambiental, 2005.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. ***Data mining: concepts and techniques***. 3. ed. Morgan Kaufmann Publishers, 2011. ISBN 978-0123814791.

HILLIER, Bill. ***Space is the machine: a configurational theory of architecture***. London: Space Syntax, 2007. Disponível em: https://discovery.ucl.ac.uk/id/eprint/3881. Acesso em: 29 mar. 2025.

HAIR, Joseph F. **Fundamentos de métodos de pesquisa em administração**. Porto Alegre: Bookman, 2005.

HAN, Jiawei; PEI, Jian; TONG, Hanghang. ***Data mining: concepts and techniques***. 4. ed. Morgan Kaufmann Publishers, 2022. ISBN 978-1892095008.

KNECHTEL, Maria do Rosário. **Metodologia da pesquisa em educação: uma abordagem teórico-prática dialogada**. Curitiba: Intersaberes, 2014. 193 p.

HOLSHEIMER, M.; APERS, P. M. G.; FLACHE, A.; ZAITSEV, A. ***A data mining perspective on databases*****.** In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, 1995. p. 150–155.

Hjørland, Birger**. What is Knowledge Organization (KO)?** ***Knowledge Organization***, v. 35, n. 2–3, p. 86–101, 2008.

JADEJISKI, Rainei Rodrigues; OLIVEIRA, Eric de; GOMES, Maurício Valeriano. **Infrações de trânsito: desdobramentos a partir do contexto Capixaba**. Revista Tocantinense de Geografia, [S. l.], v. 9, n. 18, p. 190–203, 2020. DOI: 10.20873/rtg.v9n18p190-203. Disponível em: https://periodicos.ufnt.edu.br/index.php/geografia/article/view/9706. Acesso em: 1 abr. 2025.

KNECHTEL, Maria do Rosário**. Metodologia da pesquisa em educação: uma abordagem teórico-prática dialogada.** Curitiba: Intersaberes, 2014. 193 p.

KÖCHE, José Carlos. **Fundamentos de metodologia científica: teoria da ciência e iniciação à pesquisa**. 31. ed. Petrópolis, RJ: Vozes, 2012.

LASTRES, Helena Maria Martins; ALBAGLI, Sarita; LEMOS, Cristina; LEGEY, Liz-Rejane. **Desafios e oportunidades da era do conhecimento**. São Paulo em Perspectiva, São Paulo, v. 16, n. 3, p. 60–66, 2002. Disponível em: https://www.scielo.br/j/spp/a/yHQXBsStTDwrMFHDsBhynTM/. Acesso em: 15 dez. 2024.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 8. ed. São Paulo: Atlas, 2017.

MIRANDA, Ana Maria Mendes et al. **Organização do Conhecimento e Epistemologia Social: relações teóricas, epistemológicas e aplicadas**. In: TOGNOLI, N. B.; ALBUQUERQUE, A. C.; CERVANTES, B. M. N. (org.). *Organização e representação do conhecimento em diferentes contextos: desafios e perspectivas na era da datificação*. Londrina: ISKO-Brasil: PPGCI-UEL, 2023. p. 155–177.

MOHAN, D. ***Road safety in less-motorized environments: future concerns***. International Journal of Epidemiology, Oxford, v. 31, p. 527–532, 2002.

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. ***The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*.** New York: Oxford University Press, 1997.

OPEN KNOWLEDGE BRASIL. **Índice de Dados Abertos para Cidades 2023**. Rio de Janeiro: Open Knowledge Brasil, 2024. 104 p. Disponível em: https://indicedadosabertos.ok.org.br. Acesso em: 10 abr. 2025. ISBN 978-65-993954-6-8.

PARK, J. S.; CHEN, M. S.; YU, P. S. ***An effective hash-based algorithm for mining association rules***. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, San Jose, CA, USA, 1995. p. 175–186.

RECIFE. **Portal de Dados Abertos da Prefeitura do Recife**. Disponível em: http://dados.recife.pe.gov.br/pt_BR/. Acesso em: 14 ago. 2024.

Revista Brasileira de Estudos Urbanos e Regionais. v. 11, n. 2, nov. 2009. São Paulo: ANPUR, 2009. Disponível em: https://rbeur.anpur.org.br/rbeur/issue/view/23. Acesso em: 29 mar. 2025.

ROZESTRATEN, R. J. A. **Psicologia do trânsito: conceitos e processos básicos**. São Paulo: EPU; EDUSP, 1988.

SHYALIKA, Chathurangi; WICKRAMARACHCHI, Ruwan; EL KALACH, Fadi; HARIK, Ramy; SHETH, Amit P. ***Evaluating the role of data enrichment approaches towards rare***

event analysis in manufacturing. Columbia: University of South Carolina, Artificial Intelligence Institute, 2024. Disponível em: https://scholarcommons.sc.edu/aii_fac_pub/619/. Acesso em: 14 dez. 2024.

SHERA, Jesse H. ***Sociological Foundations of Librarianship***. Bombay: Asia Publishing House, 1970.

SILVA, Fábio Henrique Vieira de Cristo e; GÜNTHER, Hartmut. **Psicologia do trânsito no Brasil: de onde veio e para onde caminha?** Temas em Psicologia, Ribeirão Preto, v. 17, n. 1, p. 121–132, 2009. Disponível em: https://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1413-389X2009000100014. Acesso em: 2 abr. 2025.

TARGINO, Maria das Graças**. Ciência da Informação e Comunicação: interconexões teórico-epistemológicas**. *Ciência da Informação*, Brasília, v. 29, n. 1, p. 7–14, jan./abr. 2000.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A**. *Data Mining: Practical Machine Learning Tools and Technique**s***. 4. ed. Amsterdam: Morgan Kaufmann, 2016.

YAGIL, D. **Drivers and traffic laws: a review of psychological theories and empirical research**. Haifa: University of Haifa, 2001.

ZAKI, M. J.; PARTHASARATHY, S.; OGDEN, W.; LIU, M. ***New algorithms for fast discovery of association rules***. *Data Mining and Knowledge Discovery*, v. 1, n. 4, p. 343–373, 1997. DOI: https://doi.org/10.1023/A:1009773317876.

ZLAPOTER, T. J. ***Determinants of motor vehicle deaths in the United States: a cross-sectional analysis***. Accident Analysis and Prevention, v. 23, p. 431–436, 1991.

APÊNDICE A – CÓDIGO PYTHON USADO PARA PROCESSAR OS DADOS

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Leitura dos dados
arquivo_csv = "registros-multas.csv"
df = pd.read_csv(arquivo_csv, delimiter=';')

df.columns = df.columns.str.strip()
df = df.rename(columns={"hora": "Horário", "registros": "Multas", "ano": "Ano"})
cores_anos = {2020: "green", 2021: "deepskyblue", 2022: "orange", 2023: "purple"}

fig, ax = plt.subplots(figsize=(12, 6))
sns.set_style("whitegrid")

for ano, cor in cores_anos.items():
    df_ano = df[df["Ano"] == ano]
    plt.plot(df_ano["Horário"], df_ano["Multas"], marker='o', linestyle='-', color=cor, label=str(ano))

# Configurações do gráfico
plt.xlabel("Horário", fontsize=12)
plt.ylabel("Registros de Multas", fontsize=12)
plt.xticks(df["Horário"].unique())

plt.legend(title="Ano", bbox_to_anchor=(1.05, 1), loc="upper right", fontsize=10)

plt.gca().set_facecolor("white")
plt.grid(False)
plt.gca().spines["top"].set_visible(False)
plt.gca().spines["right"].set_visible(False)
```

```
image_path = "grafico_multas_velocidade.png"
plt.savefig(image_path, dpi=300, bbox_inches="tight")
plt.close()
```

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Leitura dos dados
arquivo_csv = "registros-multas.csv"  # Substitua pelo nome correto do arquivo
df = pd.read_csv(arquivo_csv, delimiter=';')

df.columns = df.columns.str.strip()
df = df.rename(columns={"hora": "Horário", "registros": "Multas", "ano": "Ano"})

# Normalizar os dados por ano (Min-Max Normalization)
df["Multas_Normalizadas"] = df.groupby("Ano")["Multas"].transform(lambda x: (x - x.min())
/ (x.max() - x.min()))

cores_anos = {2020: "green", 2021: "deepskyblue", 2022: "orange", 2023: "purple"}

fig, ax = plt.subplots(figsize=(12, 6))
sns.set_style("whitegrid")

for ano, cor in cores_anos.items():
    df_ano = df[df["Ano"] == ano]
    plt.plot(df_ano["Horário"], df_ano["Multas_Normalizadas"], marker='o', linestyle='-', co-
lor=cor, label=str(ano))

plt.xlabel("Horário", fontsize=12)
plt.ylabel("Registros de Multas (Normalizadas)", fontsize=12)
plt.xticks(df["Horário"].unique())
```

```
plt.legend(title="Ano", bbox_to_anchor=(1.05, 1), loc="upper right", fontsize=10)

plt.gca().set_facecolor("white")
plt.grid(False)
plt.gca().spines["top"].set_visible(False)
plt.gca().spines["right"].set_visible(False)

image_path = "grafico_multas_normalizadas.png"
plt.savefig(image_path, dpi=300, bbox_inches="tight")
plt.close()
```

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Leitura dos dados
df = pd.read_csv("registros_meses.csv", delimiter=";")

meses = ["jan", "fev", "mar", "abr", "mai", "jun", "jul", "ago", "set", "out", "nov", "dez"]

df_norm = df.copy()

# 5. Transformar os dados para formato longo (melt) para melhor visualização
df_melted = df_norm.melt(id_vars=["ano", "hora"], value_vars=meses, var_name="mes", value_name="registros")

df_melted["ano"] = df_melted["ano"].astype("category")

cores_anos = {2020: "green", 2021: "deepskyblue", 2022: "orange", 2023: "purple"}

opacidade_meses = {
    "jan": 0.25, "fev": 0.25, "mar": 0.35, "abr": 0.35, "mai": 0.45, "jun": 0.45,
```

```
    "jul": 0.55, "ago": 0.55, "set": 0.65, "out": 0.65, "nov": 0.77, "dez": 0.75
}

# 8. Criar o gráfico
fig, ax = plt.subplots(figsize=(12, 6))
sns.set_style("whitegrid")

for (ano, mes), sub_df in df_melted.groupby(["ano", "mes"], observed=False):  # Garantir que
todas as categorias sejam consideradas
    ano = int(ano)  # Converter categoria para inteiro, caso necessário
    if ano in cores_anos:  # Garantir que só anos com cores mapeadas sejam plotados
        ax.plot(
            sub_df["hora"], sub_df["registros"], marker="o", linestyle="-",
            color=cores_anos[ano], linewidth=1.5, alpha=opacidade_meses[mes], label=f"{ano}
- {mes.upper()}"
        )

ax.set_xlabel("Horário", fontsize=12)
ax.set_ylabel("Número de Registros", fontsize=12)
ax.set_title("Distribuição das Multas por Hora e Mês", fontsize=14)

ax.set_xticks(sorted(df["hora"].unique()))

ax.set_facecolor("white")
ax.grid(False)
ax.spines["top"].set_visible(False)
ax.spines["right"].set_visible(False)
handles, labels = ax.get_legend_handles_labels()
unique_labels = dict(zip(labels, handles))
ax.legend(unique_labels.values(), unique_labels.keys(), title="Ano - Mês", bbox_to_an-
chor=(1.05, 1), loc="upper left", fontsize=9)

image_path = "distribuicao_multas_por_mes.png"
fig.savefig(image_path, dpi=300, bbox_inches="tight")
```

```
plt.close()

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Carregar os dados
df = pd.read_csv("dias_semana.csv", delimiter=";")

df_norm = df.copy()

# Definir cores específicas para cada ano
dias_cores = {2020: "green", 2021: "deepskyblue", 2022: "orange", 2023: "purple"}

fig, ax = plt.subplots(figsize=(12, 6))
sns.set_style("whitegrid")

# Plotar os dados do domingo para cada ano com sua respectiva cor
for ano in sorted(df_norm["ano"].unique()):
    df_ano = df_norm[df_norm["ano"] == ano]
    ax.plot(
        df_ano["hora"], df_ano["domingo"],
        marker="o", linestyle="-",
        color=dias_cores[ano], label=f"Domingo ({ano})",
        linewidth=2, alpha=1  # Maior espessura e opacidade total para destaque
    )

ax.set_xlabel("Horário", fontsize=12)
ax.set_ylabel("Registros de Multas", fontsize=12)
ax.set_title("Distribuição das Infrações por Hora no Domingo- Sabado", fontsize=14)
ax.set_xticks(df_norm["hora"].unique())

ax.set_facecolor("white")
ax.grid(False)
```

```
ax.spines["top"].set_visible(False)
ax.spines["right"].set_visible(False)

ax.legend()

image_path = "distribuicao_domingo_nom.png"
fig.savefig(image_path, dpi=300, bbox_inches="tight")
plt.close()
```

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Carregar os dados
df = pd.read_csv("arqujvo.csv", delimiter=";")

# Lista de dias da semana
dias_da_semana = ["segunda", "terca", "quarta", "quinta", "sexta"]

df_norm = df.copy()

cores = {
    "segunda": "orange",
    "terca": "brown",
    "quarta": "deepskyblue",
    "quinta": "purple",
    "sexta": "green",
}

# Criar o gráfico
fig, ax = plt.subplots(figsize=(12, 6))
sns.set_style("whitegrid")
```

```
legendas_ja_adicionadas = {}

for ano in df_norm["ano"].unique():
    df_ano = df_norm[df_norm["ano"] == ano]
    for dia in dias_da_semana:
        label = f"{dia.capitalize()} ({ano})"
        if label not in legendas_ja_adicionadas:
            ax.plot(
                df_ano["hora"], df_ano[dia],
                marker="o", linestyle="-",
                color=cores[dia], label=label,
                linewidth=1, alpha=1
            )

# Configuração do gráfico
ax.set_xlabel("Horário", fontsize=12)
ax.set_ylabel("Registros de Multas", fontsize=12)
ax.set_title("Distribuição das Infrações por Hora e Dia da Semana", fontsize=14)
ax.set_xticks(df_norm["hora"].unique())

ax.set_facecolor("white")
ax.grid(False)
ax.spines["top"].set_visible(False)
ax.spines["right"].set_visible(False)

ax.legend(
    title="Dia da Semana e Ano",
    loc="upper right",  # Posicionamento dentro do gráfico
    fontsize=9,
    frameon=True,

)

# Salvar a imagem
```

```
image_path = "distribuicao_dias.png"
fig.savefig(image_path, dpi=300, bbox_inches="tight")
plt.close()

fig.savefig(image_path, dpi=300, bbox_inches="tight")
plt.close()
```