UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Instituto de Biociências, Letras e Ciências Exatas- campus de São José do Rio Preto

RANIELLY APARECIDA DA SILVA

Análise Exploratória de Dados de Indicadores Criminais nos Municípios do Estado de São Paulo :

Tendências, Correlações e Desigualdades Regionais

São José do Rio Preto

2025

Ranielly Aparecida da Silva

Análise Exploratória de Dados de Indicadores Criminais nos Municípios do Estado de São Paulo :

Tendências, Correlações e Desigualdades Regionais

Dissertação, apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Mestra em Matemática.

Área de Concentração: Matemática Aplicada

Orientador: Profº Dr. Wallace Correa de Oliveira Casaca

São José do Rio Preto

2025

S586a

Silva, Ranielly Aparecida da

Análise exploratória de dados de indicadores criminais nos municípios do estado de São Paulo : tendências, correlações e desigualdades regionais / Ranielly Aparecida da Silva. -- São José do Rio Preto, 2025

62 p. : il., tabs., mapas

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Wallace Correa de Oliveira Casaca

1. Matemática aplicada. 2. Análise exploratoria. 3. Ciência de dados. 4. Criminalidade. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

RANIELLY APARECIDA DA SILVA

ANÁLISE EXPLORATÓRIA DE DADOS DE INDICADORES CRIMINAIS NOS MUNICÍPIOS DO ESTADO DE SÃO PAULO :

Tendências, Correlações e Desigualdades Regionais

Dissertação apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, para obtenção do título de Mestra em Matemática.

Área de Concentração: Matemática Aplicada

Data de defesa: 21/08/2025

BANCA EXAMINADORA

Profº Dr. Wallace Correa de Oliveira Casaca
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto

---

Profº Dr. Silvio Alexandre de Araujo
UNESP - Instituto de Biociências, Letras e Ciências Exatas - Campus de São José do Rio Preto

---

Profº Dr. Larissa Ferreira Marques
UNESP - Faculdade de Ciências - Campus de Bauru

Dedico este trabalho ao meu irmão, Robison, que sempre me incentivou, acreditou e contribuiu para um mundo mais humano. Sua luz e sua luta seguem vivas em mim.

AGRADECIMENTOS

Primeiramente agradeço a Deus e aos meus Orixás pela proteção nesse caminho, pela força que me foi dada que não me deixou desistir.

Agradeço ao meu marido, meus pais, meus tios e primos, pelo apoio e por não medirem esforços para me ajudar durante toda essa jornada.

Agradeço ao meu irmão que me incentivou a estudar e entrar no mestrado, as conversas que me deram força para seguir, e hoje, mesmo sem a sua presença física, sei que continua me dando força para realizar esse sonho, que não é só meu.

Agradeço toda a equipe do Ibilce, que possibilitou a realização desse trabalho, especialmente ao meu orientador Wallace Casaca, que sempre me acolheu e me ajudou prontamente em todos os momentos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Sem esse apoio financeiro não seria possível.

RESUMO

Esta dissertação propõe uma análise exploratória aprofundada da dinâmica criminal nos municípios do estado de São Paulo, no período compreendido entre 2017 e 2024. Reconhecendo a complexidade do fenômeno da criminalidade e sua distribuição no território paulista, o estudo busca identificar padrões espaço-temporais e investigar a relação entre as ocorrências criminais e fatores socioeconômicos. A pesquisa fundamenta-se em uma abordagem de ciência de dados, integrando e pré-processando um conjunto robusto de dados reais provenientes de fontes como a Secretaria da Segurança Pública de São Paulo e o Instituto Brasileiro de Geografia e Estatística. Os dados apresentam ocorrências mensais tipificadas em 9 classes, além de agregar informações como população estimada, área, densidade demográfica, Índice de Desenvolvimento Humano Municipal e Produto Interno Bruto dos 645 municípios do estado de São Paulo. O pré-processamento incluiu a seleção do período temporal, a integração das bases, a criação de variáveis derivadas, e a estruturação dos dados para as análises. Metodologicamente, a abordagem desenvolvida emprega estatística descritiva, análises de correlação, e análise de dados espaciais e temporais, que são representados através de tabelas, gráficos e mapas, para melhor explorar tendências, distribuições e associações das variáveis estudadas. Os resultados revelam que a criminalidade no estado de São Paulo não se distribui de forma homogênea, apresentando concentrações espaciais significativas que persistem ao longo do tempo. A análise evidencia a dependência espacial dos crimes, onde a criminalidade em um município influencia seus vizinhos, e demonstra a correlação entre os indicadores criminais e os fatores socioeconômicos, mostra ainda o efeito da pandemia COVID-19 no contexto criminal. As visualizações geradas fornecem *insights* importantes sobre a dinâmica dos crimes, facilitando a compreensão dos padrões e a identificação de áreas de maior vulnerabilidade, contribuindo, desta forma, para o aprofundamento do conhecimento sobre a criminalidade no estado de São Paulo. Os achados podem subsidiar a formulação de políticas públicas de segurança mais localizadas e eficazes, promovendo uma intervenção mais estratégica e baseada no perfil específico de cada território.

Palavras-Chave: análise exploratória; criminalidade; ciência de dados.

ABSTRACT

This dissertation proposes an in-depth exploratory analysis of criminal dynamics in municipalities in the state of São Paulo, from 2017 to 2024. Recognizing the complexity of crime occurrences and their distribution across the territory, the study identifies temporal patterns and investigates the relationship between crime occurrences and socioeconomic factors. The research is based on a data science approach, integrating and preprocessing a robust set of real data from sources such as the São Paulo Public Security Secretariat and the Brazilian Institute of Geography and Statistics. The data include monthly occurrence numbers for nine types of crimes, estimated population, area, demographic density, Human Development Index and Gross Domestic Product for the 645 municipalities in the state of São Paulo. Preprocessing included selecting the time period, integrating the databases, creating derived variables, and structuring the data for analysis. Methodologically, it employs descriptive statistics, statistical analysis, and spatial and temporal data analysis, represented through tables, graphs, and maps, to explore trends, distributions, and associations among variables. The results reveal that crime in the state of São Paulo is not distributed atmospherically, presenting significant spatial concentrations that persist over time. The analysis highlights the spatial dependence of crime, where crime in one municipality influences its neighbors, and demonstrates the interplay between crime indicators and socioeconomic factors. It also highlights the impact of the COVID-19 pandemic on crime. The visualizations obtained provide important insights into crime dynamics, facilitating the understanding of patterns and the identification of areas of greatest vulnerability. This contributes to a deeper understanding of crime in São Paulo, offering a detailed, evidence-based diagnosis. The findings support the formulation of more localized public security policies and can be effective, promoting more strategic interventions based on the specific profile of each territory.

Keywords: exploratory analysis; crime; data science.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A criminalidade é um dos desafios mais persistentes enfrentados pela sociedade brasileira. De fato, os altos índices de crimes contra a vida e patrimônio público e privado demandam estratégias eficazes de prevenção, controle e responsabilização. O estado de São Paulo, embora tenha alcançado avanços em diversos indicadores de segurança, como a redução de roubo em certas regiões, ainda apresenta uma elevada complexidade criminal, especialmente em seus grandes centros urbanos e turísticos, onde a dinâmica das ocorrências pode ser significativamente influenciada por fatores sazonais, socioeconômicos e populacionais, incluindo o impacto de grandes fluxos migratórios (FBSP, 2023), (Paz, 2023).

Neste contexto, a análise em nível municipal se torna um ponto crucial para capturar nuances que se perdem em agregações maiores, como as particularidades de municípios com economias diversificadas, infraestruturas distintas e padrões de migração sazonal específicos, que podem moldar de maneira única seus perfis criminais.

Nas últimas décadas, a crescente disponibilidade de dados públicos, aliada ao poder computacional, abriu novas fronteiras para a aplicação da ciência de dados na segurança pública. Técnicas como a Análise Exploratória de Dados (AED) para identificar *hotspots* criminais, estatística multivariada para analisar a correlação entre diferentes tipos de crime e fatores socioeconômicos, e algoritmos de aprendizado de máquina para detectar padrões complexos em grandes volumes de dados espaciais e temporais têm demonstrado seu potencial em revelar *insights* e padrões que antes eram difíceis de serem identificados.

Considerando o cenário destacado acima, a presente dissertação se insere nos campos da Matemática Aplicada (MA) e da Ciência de Dados (CD), com foco na utilização e desenvolvimento de métodos estatísticos e computacionais para compreender os intrincados padrões criminais no estado de São Paulo, a partir da coleta e avaliação de dados reais e atualizados. A aplicação de tais técnicas permite uma análise mais precisa e granular da criminalidade em nível municipal, capturando a heterogeneidade existente no território paulista.

A relevância e aplicabilidade deste trabalho reside na urgência de se adotar abordagens analíticas baseadas em evidências para compreender e enfrentar a complexa realidade da criminalidade no país. Em um cenário onde os padrões criminais são multifacetados, a utilização da CD oferece uma alternativa mais objetiva e eficiente para a geração de conhecimento a partir de metodologias do tipo *data-driven*. Ao empregar ferramentas como a AED, este estudo busca gerar percepções consistentes sobre fatores de risco específicos, padrões temporais e espaciais de crimes, além de agrupamentos de municípios com perfis criminais semelhantes, permitindo assim adotar ações preventivas direcionadas à alocação estratégica de recursos e ao desenvolvimento de políticas públicas mais efetivas para os cidadãos.

Posto os desafios e escopo da pesquisa acima, o objetivo central desta pesquisa de mestrado é analisar a evolução temporal e espacial dos crimes no estado de São Paulo visando assim

identificar padrões e tendências relevantes que possam subsidiar estratégias de prevenção e controle da criminalidade em solo paulista. Além do aspecto de mapeamento de padrões e tendências, a pesquisa explora aspectos geográficos e socioeconômicos, como tamanho da população, densidade demográfica, Produto Interno Bruto (PIB), Índice de Desenvolvimento Humano Municipal (IDHM), associados às ocorrências de 9 tipos de crimes registrados mensalmente em cada município do estado.

Desta forma, os objetivos específicos deste trabalho incluem:

- Realizar uma análise exploratória dos dados criminais, oriundos de fontes públicas oficiais, no período de 2017 a 2024 visando obter uma visão abrangente das características iniciais e da distribuição dos dados no contexto criminal.
- Identificar variações temporais e espaciais dos crimes a fim de compreender a dinâmica da criminalidade ao longo do tempo e entre os diversos contextos municipais.
- Avaliar o impacto e grau de correlação entre indicadores socioeconômicos e os índices criminais de modo a identificar e mapear possíveis associações e fatores de influência que possam explicar a variação da criminalidade.

Esta dissertação de mestrado está organizada da seguinte forma:

- O Capítulo 2 apresenta a revisão bibliográfica, contextualizando o estudo no campo da criminologia e da ciência de dados, fornecendo a fundamentação teórica basilar sobre as técnicas de análise exploratória de dados que serão empregadas, além de revisar trabalhos anteriores que são relevantes para a análise da criminalidade no estado de São Paulo e no país como um todo, identificando, desta forma, o estado atual do conhecimento e as lacunas existentes.
- O Capítulo 3 descreve detalhadamente as fontes e o tratamento dos dados utilizados na pesquisa, incluindo as variáveis criminais e socioeconômicas coletadas dos municípios paulistas ao longo do período de estudo, de 2017 a 2024, garantindo assim transparência e a replicabilidade do estudo.
- O Capítulo 4 apresenta os resultados da análise exploratória, incluindo visualizações e estatísticas descritivas, revelando os padrões iniciais nos dados criminais e socioeconômicos, tendências, sazonalidade, correlações, desigualdades regionais, fornecendo a base para as análises mais aprofundadas.
- Por fim, o Capítulo 5 discute as principais conclusões da pesquisa alinhadas com os objetivos propostos e a literatura revisada, apresenta as limitações metodológicas e sugere possíveis desdobramentos futuros para a investigação, abrindo caminho para novas pesquisas e aplicações práticas.

2 REVISÃO BIBLIOGRÁFICA

O presente capítulo dedica-se à revisão da literatura pertinente ao estudo da criminalidade focado no estado de São Paulo. Inicialmente, será apresentada uma contextualização do fenômeno criminal no âmbito brasileiro e paulista, delineando o cenário social, econômico e demográfico atual. Em seguida, a fundamentação teórica abordará os principais conceitos e as técnicas de análise exploratória de dados que serão empregadas nesta dissertação, abrangendo desde a estatística descritiva até a análise espacial e multivariada. Por fim, será realizada uma revisão de trabalhos anteriores que tratam a criminalidade focada no território paulista ou que utilizam metodologias similares, identificando assim as contribuições existentes e algumas das lacunas que a presente pesquisa busca investigar.

2.1 CONTEXTUALIZAÇÃO

Pensando no desenvolvimento da humanidade desde os primórdios, o homem, que antes vivia sozinho, passou a viver em grupos. Para o bom convívio foi preciso criar regras, mas infelizmente, nem todos as seguem. Essa quebra de regras afeta as relações sociais, sendo necessário um instrumento regulador para manter a ordem, punir os violadores e proteger a sociedade (Filho, 2023). Neste contexto surgem as leis, que são as normas que devem ser obedecidas por todos da comunidade, e infringir essas leis é o que definimos como crime. Entender o que leva o ser humano a cometer crimes, identificar se há padrões na criminalidade para tentar contê-los, é um grande desafio social, refletindo diretamente na segurança pública, sendo de fato algo que tem sido modelado e discutido em diferentes esferas, incluindo contribuições dentro da Estatística, Matemática e Ciência de Dados.

De acordo com o FBSP (2023), o Brasil apresenta elevados índices de homicídios, roubos e furtos, com variações expressivas entre regiões e municípios. O estado de São Paulo, em particular, concentra a maior metrópole do país, a cidade de São Paulo, com uma população numerosa de mais de 11 milhões de habitantes (IBGE, 2023), e uma intensa urbanização que impõe desafios específicos à segurança pública. Sendo um polo industrial, financeiro e de serviços diversificado, a capital paulista apresenta uma complexa dinâmica socioeconômica que pode influenciar os padrões e a natureza da criminalidade em suas diferentes regiões.

Apesar de registrar redução em alguns indicadores criminais nas últimas décadas, como por exemplo, uma queda de 55.3% na taxa de homicídios entre os anos de 2012 e 2022 (FBSP, 2023), o estado de São Paulo continua apresentando desafios como altos índices de criminalidade em municípios litorâneos independentemente da sazonalidade turística (Paz, 2023). Adicionalmente, o estado caracteriza-se por uma acentuada diversidade socioeconômica entre suas regiões, com áreas de elevado desenvolvimento contrastando com municípios que enfrentam maiores vulnerabilidades sociais. Essa heterogeneidade, refletida em indicadores como o Ín-

dice de Desenvolvimento Humano Municipal (IDHM), a distribuição de renda, e os níveis de escolaridade, pode estar intrinsecamente ligada à distribuição e aos tipos de criminalidade observados em diferentes localidades do estado. Nos municípios litorâneos, por exemplo, a dinâmica socioeconômica, influenciada pela sazonalidade do turismo, pode gerar flutuações populacionais e alterações nas oportunidades econômicas que, por sua vez, podem impactar os índices de criminalidade de maneiras complexas, nem sempre seguindo diretamente os picos da atividade turística.

A crescente digitalização dos registros policiais, combinada com a disponibilidade de bases de dados abertas têm permitido o emprego de métodos analíticos para compreender o fenômeno da criminalidade com maior profundidade. Nesse contexto, modelos matemáticos e técnicas do tipo *data-driven* têm se mostrado ferramentas poderosas, que combinam métodos numéricos, estatísticos e computacionais para extrair conhecimento a partir de grandes volumes de dados (Provost; Fawcett, 2013). Seu uso tem se expandido para diversas áreas, incluindo a criminologia, onde a análise baseada em dados pode revelar padrões ocultos, antecipar tendências e apoiar decisões estratégicas relacionadas à segurança pública.

A ciência de dados permite a análise e interpretação de grandes quantidades de registros criminais, corriqueiramente estruturados de forma temporal e geográfica. Dentre as etapas fundamentais nesse processo, destaca-se a Análise Exploratória de Dados (AED), que oferece uma visão ampla sobre as características do conjunto de dados.

2.2 FUNDAMENTAÇÃO TEÓRICA

A AED é uma abordagem computacional e estatística fundamentalmente descritiva e geradora de hipóteses, que visa identificar padrões, tendências, detectar anomalias, e sugerir relações, utilizando técnicas visuais e quantitativas para resumir as principais características de um conjunto de dados. Essa abordagem foi amplamente popularizada e formalizada pelo estatístico Tukey (1977) em sua obra "*Exploratory Data Analysis*", que enfatiza a importância de examinar os dados de forma flexível e visual para gerar *insights* iniciais. Courtney (2021) descreveu a AED como um processo iterativo e aberto, permitindo que os analistas examinem os dados sem preconceitos pré-estabelecidos, facilitando a compreensão profunda das informações disponíveis e guiando as etapas subsequentes da investigação.

A AED é essencial para entender a estrutura e o comportamento dos dados, sendo aplicada de forma iterativa para detectar padrões, *outliers*, *hotspots*, relações entre variáveis e características relevantes, além do uso extensivo de visualizações gráficas que podem gerar novas perguntas (Courtney, 2021). A AED é um importante ponto de partida para qualquer análise estatística ou modelagem, fornecendo a base para responder às questões de pesquisa.

Na criminologia, a AED possibilita a identificação de padrões espaciais e temporais em registros criminais, auxiliando na compreensão das dinâmicas da criminalidade e na formulação de hipóteses para análises mais aprofundadas. Para a realização desta etapa dentro do escopo de

nossa pesquisa, será utilizada a linguagem de programação *Python*, que oferece um ecossistema robusto de bibliotecas voltadas para a análise de dados. Entre as principais bibliotecas, segundo McKinney (2018), destacam-se:

- *Pandas*: possui diferentes ferramentas e métodos voltados para análises capazes tanto de esmiuçar os dados em nível mais básico como realizar operações mais complexas.
- *NumPy*: é adequada para a realização de cálculos numéricos e científicos, oferece vetorização de funções matemáticas.
- *Matplotlib*: permite a visualização de dados sob diferentes perspectivas, incluindo a confecção de gráficos de linhas, histogramas, entre outros.
- *Seaborn*: baseada na biblioteca *Matplotlib*, facilita a criação de visualizações mais complexas como mapas de calor e *Violin Plot*.

Essas ferramentas serão empregadas para viabilizar o uso de técnicas de AED, tais como: estatísticas descritivas, visualizações gráficas, análise de correlação, análise de séries temporais, análise de dados espaciais e análise multivariada de dados, com o objetivo de investigar a evolução temporal e a distribuição geográfica das ocorrências criminais no estado de São Paulo, bem como para explorar a relação entre variáveis socioeconômicas e demográficas, como população, densidade demográfica e IDHM, e os índices de criminalidade nos municípios paulistas entre 2017 e 2024. A seguir, algumas dessas técnicas de AED serão discutidas em detalhes.

2.2.1 Estatística descritiva

A estatística descritiva fornece um resumo quantitativo das principais características dos dados de maneira concisa e informativa, sendo fundamental para a etapa inicial de compreensão do conjunto de dados criminais e socioeconômicos que serão analisados nesta dissertação. As medidas utilizadas incluem:

- Média e mediana: ambas as medidas estatísticas indicam a tendência central dos dados. A mediana é o elemento central dos dados ordenados e é menos sensível a valores extremos. Já a média, calculada pela soma de todos os valores dividida pelo número total de dados, representa o valor central esperado e, ao ser comparada com a mediana, pode revelar a assimetria na distribuição das taxas de criminalidade ou dos indicadores socioeconômicos entre os municípios (Feijoo, 2010).
- Desvio padrão, variância e amplitude: estas medidas quantificam a variabilidade dos dados e identificam a dispersão dos valores. Segundo Bastos & Duquia (2007), a amplitude, obtida pela diferença entre o valor máximo e o mínimo, oferece uma visão geral da dispersão. A variância considera todos os valores através do somatório do quadrado da distância de cada valor em relação à média, e o desvio padrão, que é a raiz quadrada da

variância, estimam o quanto cada valor se distancia da média aritmética. Um alto desvio padrão pode indicar uma significativa heterogeneidade nos dados.

- Quartis: são subintervalos que auxiliam na compreensão da distribuição e na identificação de valores extremos, dividindo os dados ordenados em quatro partes iguais. O primeiro quartil (Q1) representa o intervalo que concentra os primeiros 25% dos dados, a mediana (segundo quartil, Q2) os 50%, e o terceiro quartil (Q3) os 75% (Feijoo, 2010). Considerando o contexto de nosso estudo, a análise dos quartis permite identificar a concentração das taxas de criminalidade em diferentes faixas de municípios, podendo sinalizar a presença de potenciais *outliers* nos extremos da distribuição.

A estatística descritiva, portanto, fornece um panorama inicial das características quantitativas dos dados criminais e socioeconômicos dos municípios paulistas. No entanto, a visualização de dados complementa essa análise, revelando padrões, distribuições e relações de forma gráfica, facilitando a identificação de tendências e anomalias que podem não ser evidentes apenas em tabelas e medidas resumidas (Cleveland, 1993), (Tukey, 1977). A seção seguinte detalhará as diversas técnicas de visualização que serão empregadas nesta dissertação.

2.2.2 Visualização de dados

A visualização de dados é uma ferramenta essencial presente no contexto da AED, uma vez que permite a representação de informações complexas de forma acessível e intuitiva e auxiliando na descoberta de padrões e tendências que podem não ser evidentes em análises puramente quantitativas.

Para estabelecer uma visualização adequada dos dados, precisamos alinhar alguns pontos como ter um conjunto de dados "limpo", representado em formato apropriado para a ferramenta de visualização, ter uma única mensagem (pergunta chave) que será destacada, escolher a forma de visualização de acordo com a pergunta específica que se busca responder, destacar o que é essencial e aplicar princípios de *design* que garantam clareza e evitem distrações (Silva, 2019).

No âmbito da análise de dados criminais, diversas técnicas podem ser aplicadas para revelar padrões espaciais, temporais e relacionais como:

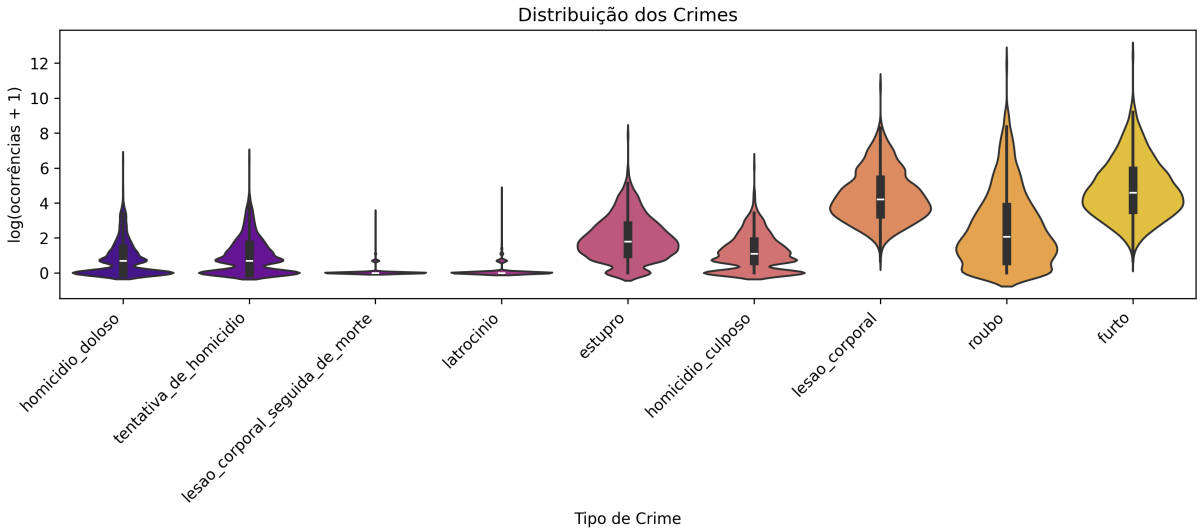
- Tabelas: organiza dados de forma estruturada em linhas e colunas, possibilita uma visão global, facilita a leitura de valores precisos e a comparação direta entre diferentes categorias ou variáveis (Falco, 2008).

- Gráficos de Linha: usados para representar séries temporais, são essenciais para visualizar a evolução temporal do número de ocorrências de cada tipo de crime ao longo dos meses e anos em nível municipal e estadual, permitem identificar tendências de aumento, diminuição ou sazonalidade (Falco, 2008).

- *Boxplot*: apresenta valores de tendência central, dispersão e simetria dos dados agrupados, facilita a comparação da distribuição entre diferentes tipos de crimes ou regiões. É composto pelos limites inferior e superior nas hastes e pela caixa, que é definida através dos valores dos quartis, Q1, mediana (Q2) e Q3. Quanto menor a caixa, mais concentrado em torno da mediana os dados estão. O limite inferior corresponde até $Q1 - 1,5 \times IQR$, e o superior é até $Q3 + 1,5 \times IQR$, onde IQR é o intervalo interquartil ($Q3 - Q1$). Se existir valores abaixo do limite inferior e/ou acima do limite superior temos os *outliers*. *Boxplot* é um dos métodos que encontra *outliers*, essa identificação é importante pois, valores extremos podem levar as análises para resultados incertos (Tukey, 1977).

- *Violin Plot*: é uma representação gráfica introduzida por Hintze & Nelson (1998), que usa a densidade de probabilidade e as características do *boxplot*, criando uma curva (parecida com um violino) que mostra a distribuição dos dados. A Figura 1 exibe um exemplo de *Violin Plot* mostrando a distribuição média de 9 tipos de crimes no estado de São Paulo de 2017 a 2024, usando escala logarítmica.

Figura 1 – Exemplo de *Violin Plot*



Fonte: Elaborada pela autora

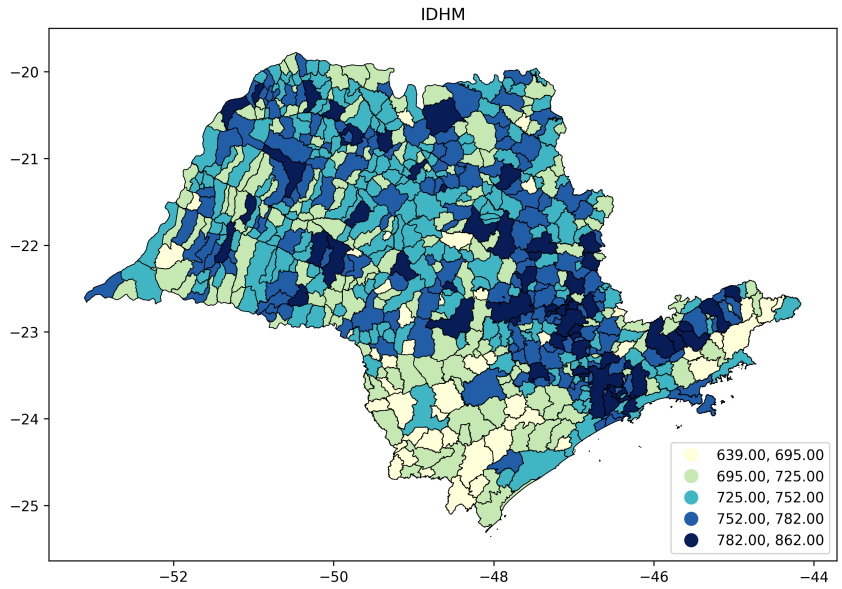
- Histogramas: são úteis, por exemplo, para analisar a distribuição das taxas de criminalidade por município. Os dados são divididos igualmente em intervalos e cada intervalo possui uma frequência que é representada por retângulos (Falco, 2008).

- Gráficos de Dispersão: permitem explorar a relação entre duas variáveis, como por exemplo, a taxa de roubo e densidade demográfica dos municípios (Cleveland, 1993). Ele representa cada observação como um ponto em um plano cartesiano, onde a posição do ponto é determinada pelos valores das duas variáveis. Se existe uma boa correlação entre as variáveis os pontos se acumulam em uma linha imaginária, podendo ser uma correlação positiva, com pontos na diagonal crescente para a direita, ou negativa, na

diagonal crescente para esquerda, e quando os pontos ficam dispersos a correlação é fraca ou nula.

- Mapas Coropléticos: são representações visuais onde as características de áreas geográficas são codificadas por cores para comunicar a distribuição espacial de uma variável quantitativa agregada (Andrienko; Andrienko; Savinov, 2001). São importantes, por exemplo, para visualizar a distribuição espacial da criminalidade no estado de São Paulo. Ao colorir os municípios de acordo com a intensidade das taxas de criminalidade é possível identificar *hotspots* (áreas de alta criminalidade) e outros padrões geográficos na desigualdade da criminalidade. A Figura 2 mostra um exemplo de mapa coroplético que exibe a distribuição do IDHM em cada município do estado de São Paulo.

Figura 2 – Exemplo de mapa coroplético



Fonte: Elaborada pela autora

- Mapas de Calor: utilizam variações de cor para representar a magnitude dos valores em uma matriz de dados, sendo úteis para visualizar correlações entre variáveis ou padrões de valores em múltiplas observações (Wilkinson, 2005).

- Gráficos de Barras: adequados para comparar, por exemplo, o número total de ocorrências de diferentes tipos de crimes em um determinado período ou ainda comparar as taxas de criminalidade entre diferentes grupos de municípios. É a representação dos dados através de retângulos, dispostos em colunas, onde cada coluna é uma variável e a altura é o valor que cada variável possui (Falco, 2008).

A escolha das ferramentas de visualização vai de acordo com o que se deseja mostrar, a seguir veremos cada tipo de análise, e quais ferramentas podem ser utilizadas.

2.2.3 Análise de correlação

A análise de correlação em dados criminais procura quantificar e descrever a força e a direção da relação linear entre duas ou mais variáveis relacionadas a fenômenos criminais ou a fatores que podem influenciá-los. Ao aplicar essa técnica ao estudo da criminalidade no estado de São Paulo, por exemplo, podemos investigar como as taxas de criminalidade se associam às variáveis socioeconômicas ou a outros crimes. Para determinar essa correlação, podemos utilizar:

- Correlação de Pearson: mede a força e a direção de uma relação linear entre duas variáveis quantitativas. O coeficiente de correlação (r) varia de -1 a 1, onde valores próximos a 1 indicam uma correlação positiva, valores próximos a -1 indicam uma correlação negativa e valores próximos a 0 indicam uma correlação fraca ou inexistente (Mukaka, 2012). Vale salientar que uma alta correlação entre variáveis não define necessariamente uma causalidade entre elas. Além disso, a correlação de Pearson mede relações lineares, podendo não capturar associações não lineares importantes. A fórmula da correlação de Pearson é dada por:

  $$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum(x_i - \overline{x})^2)(\sum((y_i - \overline{y})^2)}}$$

  onde $\overline{x}$ é a média dos valores x_i, da variável x e $\overline{y}$ é a média dos valores y_i da variável y.

- Correlação de Spearman: é uma medida não paramétrica da associação estatística entre duas variáveis. Ao contrário da correlação de Pearson, que avalia a relação linear entre duas variáveis quantitativas assumindo distribuição normal, a correlação de Spearman avalia a monotonicidade da relação, sendo apropriada quando os dados não atendem aos pressupostos de normalidade ou contêm *outliers* que poderiam distorcer o resultado da correlação linear (Mukaka, 2012).

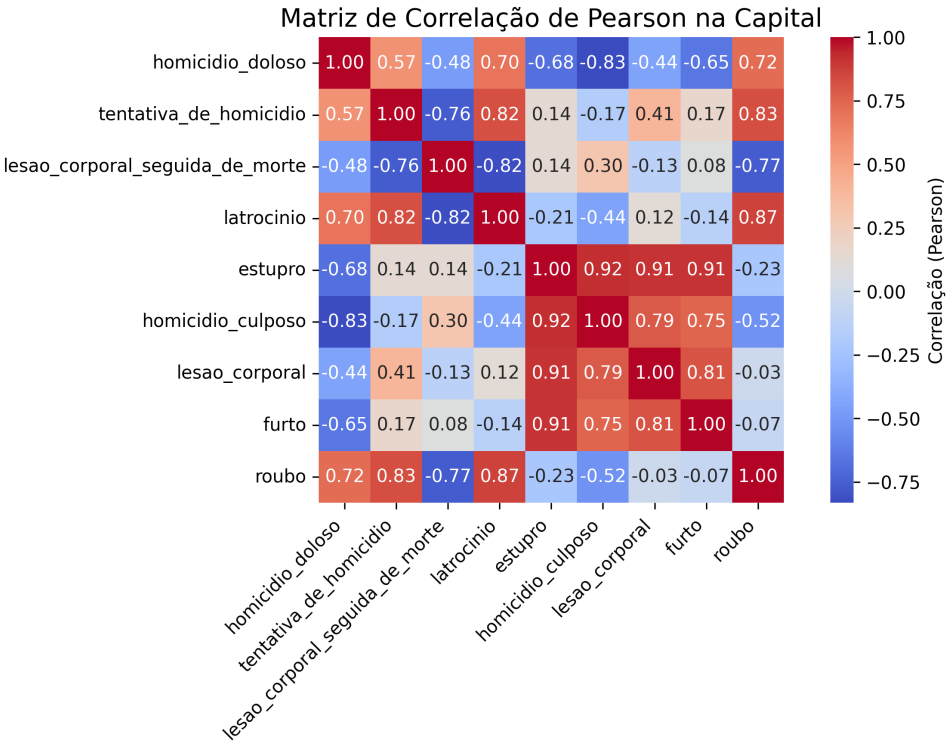
  O coeficiente de Spearman é calculado a partir dos postos (*ranks*) das observações, ou seja, os valores das variáveis são ordenados, e a correlação é calculada com base nessas posições. Isso o torna robusto frente a escalas não lineares e valores discrepantes, sendo indicado especialmente em estudos com variáveis ordinais, assimétricas ou em amostras pequenas (Conover, 1999). Seguindo a fórmula:

  $$\rho = 1 - \frac{\sum_{i=1}^{n} d^2}{n(n^2 - 1)}$$

  onde ρ é o coeficiente de Spearman, d é a distância dos postos, n é o número de variáveis. O valor de ρ varia entre -1 e 1; ρ = 1: correlação positiva perfeita (à medida que uma variável aumenta, a outra também aumenta); $\rho = -1$: correlação negativa perfeita; ρ = 0: ausência de relação entre as variáveis.

- Matriz de Correlação: apresenta os coeficientes de correlação para todos os pares de variáveis, para analisar a relação entre múltiplas variáveis criminais e/ou socioeconômicas simultaneamente, possibilitando identificar padrões de relação dentro do conjunto de dados. Essa matriz é frequentemente visualizada através de mapas de calor, que facilitam a identificação rápida de correlações fortes e fracas entre as variáveis. A Figura 3 mostra uma matriz de correlações entre dados criminais na capital paulista, representada em um mapa de calor.

Figura 3 – Exemplo de matriz de correlação em um mapa de calor



Fonte: Elaborada pela autora

A identificação de correlações significativas pode direcionar investigações futuras e auxiliar na formulação de hipóteses sobre os fatores que podem estar associados aos níveis de criminalidade nos municípios de São Paulo.

Enquanto a análise de correlação examina as relações estáticas entre variáveis, a análise de séries temporais direciona o foco para a dimensão temporal dos dados de ocorrências criminais. Ao analisar as sequências de dados coletados ao longo do tempo, é possível identificar tendências, padrões de sazonalidade, elementos necessários para compreender a dinâmica temporal da criminalidade (Morettin; Toloi, 2017).

2.2.4 Análise de séries temporais

Uma sequência de dados ordenados cronologicamente constitui uma série temporal. A análise da dimensão temporal da criminalidade é fundamental para compreender sua dinâmica

evolutiva, pois possibilita modelar e entender os padrões subjacentes nesses dados, permitindo a identificação de tendências, sazonalidade e outros componentes que influenciam a criminalidade ao longo do tempo (Hyndman; Athanasopoulos, 2021). Uma série temporal de ocorrências criminais pode ser decomposta em:

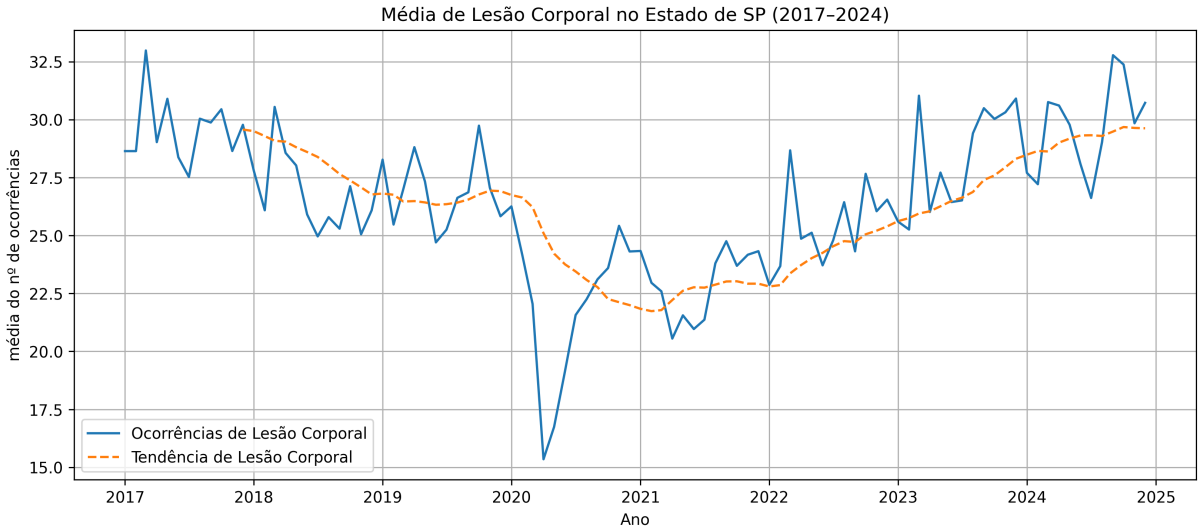
- Tendência: o movimento de longo prazo da série. Pode ser crescente, decrescente ou estável ao longo dos anos, refletindo mudanças socioeconômicas ou a implementação de políticas de segurança de longo prazo. Identificar a tendência ajuda a entender a direção geral da criminalidade (Hyndman; Athanasopoulos, 2021).
- Sazonalidade: são padrões que se repetem em intervalos fixos e conhecidos, geralmente dentro de um ano (mensal, trimestral). Na criminalidade, pode haver picos em certos meses devido a férias, feriados ou eventos específicos, como um aumento de furtos em períodos de maior circulação de pessoas ou picos de violência associados a celebrações (Hyndman; Athanasopoulos, 2021).
- Ciclos: são flutuações de longo prazo que não têm um período fixo e estão relacionadas a fatores econômicos ou sociais mais amplos (Hyndman; Athanasopoulos, 2021).
- Ruído (Resíduo ou Componente Irregular): são as variações aleatórias e imprevisíveis na série que não podem ser atribuídas aos outros componentes.

A fim de determinar esses padrões, são necessárias algumas técnicas de análise de séries temporais, tais como:

- Gráficos de linhas: uma das principais técnicas é a visualização por meio de gráficos. Plotar as ocorrências criminais ao longo do tempo permite uma inspeção visual inicial da tendência e sazonalidade (Cleveland, 1993). A Figura 4 ilustra esse tipo de visualização, apresentando a média mensal de ocorrências de lesão corporal no estado de São Paulo entre 2017 e 2024. A linha azul mostra a variação mensal da média de casos enquanto a linha tracejada laranja evidencia a tendência geral da série.
- Decomposição da série temporal: técnicas como a decomposição aditiva ou multiplicativa, usadas para separar os diferentes componentes da série como tendência, sazonalidade e resíduo, facilitando assim a análise individual de cada um (Hyndman; Athanasopoulos, 2021). Na decomposição aditiva, a tendência pode ser estimada por uma média móvel centralizada, que suaviza variações curtas ao longo do tempo, destacando o comportamento de longo prazo da série.

  No caso do estudo em questão, a média móvel centralizada considerou a média dos valores do mês atual, em adição aos valores das médias dos 6 meses anteriores, 5 meses posteriores, e assim, gradativamente, para todos os meses em análise. A sazonalidade foi obtida subtraindo-se a tendência da série original e calculando a média dos valores

Figura 4 – Exemplo de gráfico de linhas para análise temporal



Fonte: Elaborada pela autora

restantes para cada mês, refletindo padrões sistemáticos que se repetem em ciclos regulares (Hyndman; Athanasopoulos, 2021). O resíduo corresponde à diferença entre a série original e a soma da tendência e da sazonalidade, representando variações aleatórias não explicadas pelos demais componentes.

A análise da criminalidade não se limita apenas a sua evolução temporal, mas também à sua distribuição geográfica. A análise de dados espaciais explora os padrões das ocorrências criminais entre os municípios de São Paulo, investigando a presença de autocorrelação e a formação de *hotspots* (Anselin, 1988) (Getis; Ord, 1992). A seguir, é apresentado a fundamentação sobre as técnicas mais populares desta frente.

2.2.5 Análise exploratória de dados espaciais

A Análise Exploratória de Dados Espaciais (AEDE) é um ramo da estatística dedicado a descrever a distribuição espacial dos dados, encontrar padrões de dependência e detectar formas de heterogeneidade espacial (Anselin, 2019). Por exemplo, diferentemente das abordagens estatísticas tradicionais, a AEDE reconhece explicitamente que os fenômenos criminais não ocorrem de forma aleatória no espaço, e que a localização geográfica pode influenciar significativamente a ocorrência dos crimes. No contexto em questão, a AEDE se torna uma ferramenta valiosa para investigar se a ocorrência de crimes tende a se aglomerar em certas regiões do estado, se há áreas com níveis atipicamente altos ou baixos de criminalidade e se fatores espaciais podem estar relacionados à distribuição dos delitos.

A dependência espacial (ou autocorrelação espacial) refere-se ao grau em que valores de uma variável em locais próximos são semelhantes (autocorrelação espacial positiva) ou diferentes (autocorrelação espacial negativa) do que seriam esperados por acaso (Goodchild, 1986). Considerando o contexto desta pesquisa, a presença de autocorrelação espacial positiva sugere que a

taxa de criminalidade em um município pode estar sendo influenciada pela criminalidade em seus vizinhos, possivelmente devido a processos de contágio espacial ou a influência de fatores contextuais compartilhados entre áreas próximas.

Já a heterogeneidade espacial reflete a variação não uniforme de um fenômeno ao longo do espaço. Na análise de criminalidade, isso pode se manifestar na existência de *hotspots* (aglomerações de municípios com alta criminalidade) e *coldspots* (aglomerações de municípios com baixa criminalidade).

Algumas técnicas de AEDE consistem em:

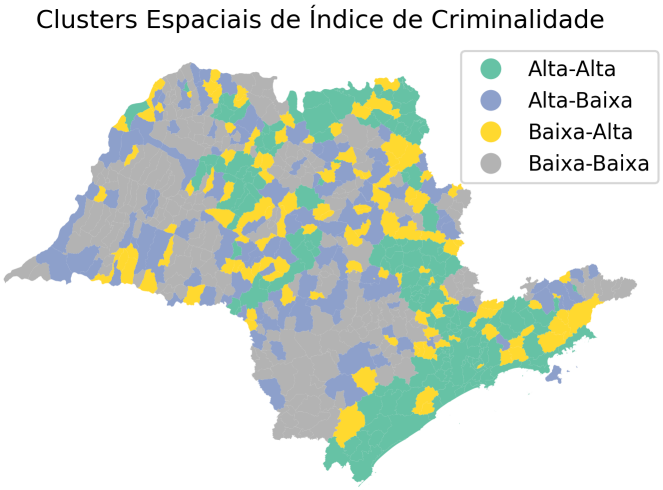
- Visualização Espacial: mapas temáticos, como os mapas coropléticos e mapas de calor, são as bases da visualização espacial, permitindo identificar visualmente padrões na distribuição da criminalidade entre os municípios (Andrienko; Andrienko; Savinov, 2001).

- Estatísticas Globais de Autocorrelação Espacial: medem, segundo Anselin (1988), o grau geral de autocorrelação espacial para toda a área de estudo. A medida mais usada é o Índice de Moran Global que varia de -1 a 1, onde valores próximos de 1 indicam autocorrelação espacial positiva (aglomeração), valores próximos de -1 indicam autocorrelação espacial negativa (dispersão), e valores próximos a zero indicam ausência de autocorrelação espacial (Moran, 1950).

- Estatísticas Locais de Autocorrelação Espacial (LISA): segundo Anselin (2019), permitem identificar *clusters* (aglomerações) de valores semelhantes, conforme a seguinte taxonomia: i) Alta-Alta – municípios com alta taxa criminal estão agrupados espacialmente próximos a outros municípios que também têm alta taxa criminal; ii) Baixa-Baixa – municípios com baixa taxa estão agrupados próximos a outros municípios com taxa baixa; e *outliers* espaciais: iii) Alta-Baixa – municípios com alta taxa estão agrupados próximos a outros municípios com taxa baixa; iv) Baixa-Alta – municípios com taxa baixa estão agrupados próximos a outros municípios com alta taxa. Para essas medidas, são usadas o Índice de Moran Local que é uma decomposição do Índice de Moran Global que avalia a autocorrelação espacial para cada unidade geográfica individual (Anselin, 1995). Os resultados podem ser visualizados em mapas de *clusters*, identificando os tipos de aglomeração: *hotspots*, *coldspots* e *outliers* espaciais. A Figura 5 a seguir apresenta um exemplo de mapa do estado de São Paulo representando os *clusters* de Índice de Criminalidade de seus municípios.

Ao aplicar a AEDE aos dados de criminalidade por município, conseguimos identificar se há padrões de aglomeração espacial para os diferentes tipos de crimes e para o índice de criminalidade. É possível mapear os *hotspots* e *coldspots* de criminalidade, revelando áreas com concentrações significativamente altas ou baixas de delitos, o que pode auxiliar na identificação de áreas prioritárias para intervenção e estudo mais aprofundado, fornecendo respostas à questões sobre a distribuição geográfica da criminalidade. Podemos detectar *outliers* espaciais, municípios

Figura 5 – Exemplo de mapa de *clusters*



Fonte: Elaborada pela autora

com taxas de criminalidade atipicamente altas ou baixas em relação aos seus vizinhos. Esses *outliers* podem indicar fatores locais específicos que influenciam a criminalidade. E ainda, explorar a relação espacial entre a criminalidade e variáveis socioeconômicas, investigando se municípios com características socioeconômicas semelhantes tendem a ter níveis de criminalidade semelhantes, considerando também a autocorrelação espacial das variáveis socioeconômicas.

Para obter uma compreensão mais holística da interação entre múltiplos fatores relacionados à criminalidade, a análise multivariada de dados examina ao mesmo tempo as relações entre diversas variáveis. Essa abordagem permite identificar padrões mais complexos (Hair et al., 2009), (Tabachnick; Fidell, 2019).

2.2.6 Análise multivariada

A Análise Multivariada Exploratória (AME) compreende um conjunto de técnicas estatísticas e gráficas que visam explorar a estrutura de dados com múltiplas variáveis simultaneamente, sem a imposição de modelos causais predefinidos (Hair et al., 2009). Diante do desafio de analisar a complexa interação entre diversos fatores relacionados à criminalidade, a AME se torna uma ferramenta valiosa para descobrir padrões e gerar hipóteses sobre essas relações. Diferentemente da análise univariada, que examina variáveis isoladamente e da análise bivariada, que investiga a relação entre dois pares de variáveis, a AME busca identificar padrões, relações e estruturas subjacentes em conjuntos de dados complexos, facilitando a geração de hipóteses e a compreensão das interações entre as variáveis (Tabachnick; Fidell, 2019).

No contexto da análise exploratória de dados criminais, a AME pode ser particularmente útil para investigar como diferentes tipos de crimes co-ocorrem, identificar grupos de municípios com perfis de criminalidade semelhantes com base em múltiplas taxas de crimes e variáveis socioeconômicas, e reduzir a dimensionalidade de um conjunto de dados complexo para facilitar

a visualização e a interpretação.

As principais técnicas utilizadas para realizar a AME são:

- Redução de Dimensionalidade: quando o número de variáveis é elevado, técnicas de redução de dimensionalidade podem simplificar a análise e a visualização, preservando a maior parte da variabilidade dos dados. A técnica mais comum é a Análise de Componentes Principais (PCA): que transforma um conjunto de variáveis correlacionadas em um número menor de variáveis não correlacionadas (componentes principais) que capturam a maior parte da variância dos dados originais (Jolliffe; Cadima, 2016). A interpretação desses componentes principais, baseada no conhecimento do contexto da criminalidade e dos fatores socioeconômicos, é fundamental. A PCA pode revelar as dimensões mais importantes que explicam a variabilidade nos perfis de criminalidade e socioeconômicos dos municípios.

- Análise de Agrupamentos: busca identificar grupos homogêneos de observações, com base na similaridade de seus valores em múltiplas variáveis. Algoritmos de *clustering* como *k-means* podem revelar, por exemplo, agrupamentos de municípios com padrões de criminalidade ou características socioeconômicas semelhantes, sendo o número de *clusters* guiada por critérios estatísticos e pela interpretabilidade no contexto da pesquisa (Everitt et al., 2011).

- Estratégias de Visualização: a representação gráfica desses dados pode ser feita por meio de gráficos de dispersão, mapas de calor e mapas coropléticos.

A Análise Multivariada Exploratória pode ser aplicada nos dados criminais para identificar perfis de criminalidade, utilizando *clustering* para agrupar municípios com padrões semelhantes, conforme as taxas dos 9 tipos de crimes que foram explorados neste estudo. Neste caso, pode-se explorar a relação entre criminalidade e fatores socioeconômicos, aplicando a técnica PCA para identificar as dimensões socioeconômicas mais importantes que explicam a variabilidade nas taxas de criminalidade entre os municípios. Em suma, a AME fornece uma compreensão mais aprofundada sobre a relação entre os diversos aspectos da criminalidade e os fatores socioeconômicos no estado de São Paulo.

2.2.7 Aprendizado de máquina

Em adição às abordagens de natureza exploratória, vale ressaltar a importância da frente de aprendizado de máquina, especialmente no contexto de uma base de dados de alta dimensionalidade, uma vez que tal área oferece diferentes algoritmos capazes de automatizar a descoberta de padrões, identificar relações não lineares, e agrupar observações de maneira potencialmente mais eficiente e escalável.

O aprendizado de máquina tem se tornado uma ferramenta cada vez mais utilizada na análise de dados complexos, incluindo dados sociais e criminais, pois seus algoritmos podem oferecer

abordagens alternativas e mais sofisticadas para tarefas que também são realizadas na AED tradicional, além de poder revelar padrões não lineares e interações complexas que podem passar despercebidos em análises exploratórias mais convencionais.

Tendo estabelecido a estrutura teórica das técnicas de Ciência de Dados que serão empregadas, a seção subsequente direcionará o foco para a revisão de trabalhos anteriores que investigaram a criminalidade no estado de São Paulo ou em contextos similares, identificando as abordagens metodológicas já empregadas, os principais achados e as lacunas que o presente estudo busca preencher.

2.3 REVISÃO DE TRABALHOS ANTERIORES

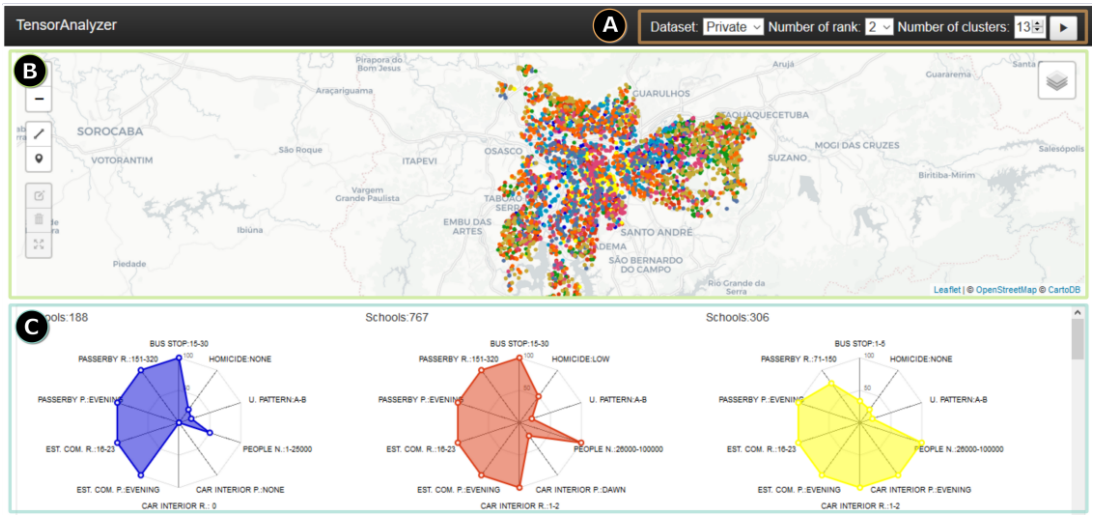
A presente seção tem como objetivo apresentar e analisar estudos anteriores relevantes para a compreensão da criminalidade, especialmente no contexto do estado de São Paulo, que dialogam com a aplicação de métodos de Ciência de Dados e que consideram as dimensões espacial, temporal e socioeconômica, pilares da presente dissertação. A revisão busca identificar as abordagens metodológicas já empregadas, os principais achados e as lacunas que o presente estudo pretende abordar, fornecendo o contexto para a investigação da dinâmica criminal nos municípios paulistas.

Um número significativo de trabalhos recentes tem explorado a dinâmica da criminalidade em grandes centros urbanos, incluindo a metrópole São Paulo, através da lente da análise espaço-temporal e do poder das ferramentas de visualização. Neste prisma, Nery et al. (2022) introduziram uma abordagem inovadora com a técnica de Fatoração de Tensor Não Negativa (NTF) no *framework* computacional denominado *TensorAnalyzer*, permitindo a identificação de padrões complexos que se manifestam simultaneamente no espaço, no tempo e em diferentes categorias de crimes, Figura 6. Essa metodologia revela agrupamentos de áreas com comportamentos criminais similares ao longo do tempo e a detecção de anomalias.

Em uma linha complementar, a tese de Zanabria (2021) e o artigo (Garcia et al., 2021), que originou a ferramenta analítica assistida por visualização denominada *CrimAnalyzer*, demonstram a eficácia da combinação de *visual analytics*, mineração de dados e aprendizado de máquina para a exploração de padrões espaço-temporais em dados criminais paulistanos, Figura 7. Essas ferramentas facilitam a identificação de *hotspots*, a análise da evolução temporal das ocorrências, e a comparação entre diferentes regiões, com interfaces projetadas para serem acessíveis a tomadores de decisão.

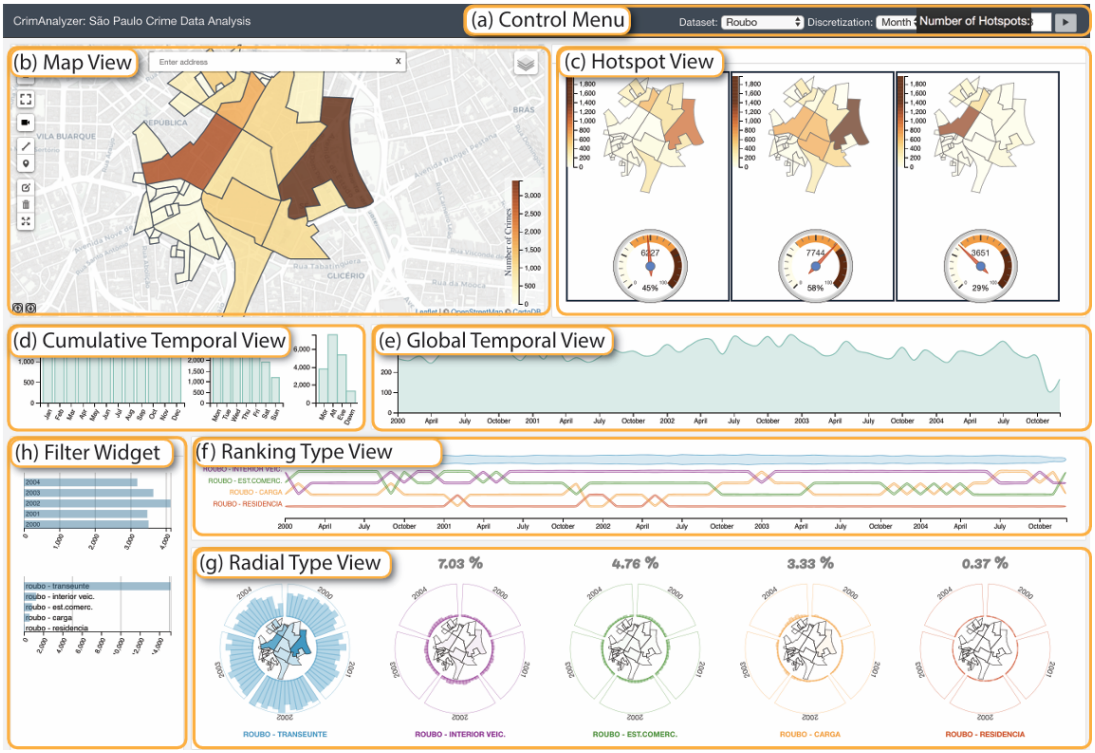
Avançando nessa direção, Zanabria et al. (2022) apresentaram o *CriPAV*: um sistema que analisa padrões criminais em nível de rua, considerando tanto a intensidade quanto a probabilidade de ocorrência dos crimes, além da similaridade de padrões temporais entre locais distintos. Em comum, esses estudos enfatizam a natureza dinâmica e espacialmente concentrada da criminalidade, bem como o potencial das técnicas de visualização interativa para revelar padrões complexos e fornecer *insights* valiosos para a segurança pública. Embora empreguem diferentes

Figura 6 – Sistema *TensorAnalyzer*: a visualização de padrões permite a compreensão da relação entre crimes e outras variáveis envolvidas na análise. Nossa ferramenta visual compreende um Menu de Controle (A), Visualização de Mapa (B) e Visualização de Padrões (C).



Fonte: (Nery et al., 2022)

Figura 7 – Sistema *CrimAnalyzer*: as visualizações interativas espaciais e temporais permitem a exploração de regiões locais e revelam seus padrões criminais ao longo do tempo



Fonte: (Garcia et al., 2021)

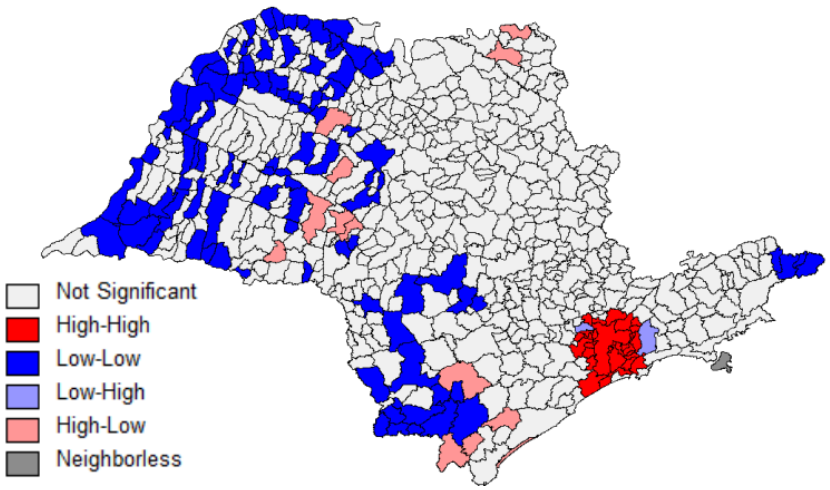
metodologias de análise espaço-temporal e visualização, todos convergem na importância de ir além das análises agregadas, explorando as nuances da distribuição criminal ao longo do tempo e no espaço, uma perspectiva central para a presente dissertação na análise dos padrões criminais

nos municípios do estado de São Paulo.

A crescente aplicação de aprendizado de máquina na análise de padrões criminais espaço-temporais também é evidenciada na revisão sistemática de Butt et al. (2020). Os autores destacam a eficácia de algoritmos como o *DBSCAN* na detecção de *hotspots*, ao mesmo tempo, apontam desafios como a disponibilidade limitada de dados públicos integrando informações espaciais, temporais e socioeconômicas. Esses desafios ressoam no contexto paulista, onde a heterogeneidade municipal exige abordagens analíticas adaptadas às especificidades locais.

A importância de incorporar a dimensão espacial na análise da criminalidade no estado de São Paulo já era um ponto evidenciada em estudos anteriores. O trabalho de Almeida (2007) investigou os padrões de distribuição e a presença de autocorrelação espacial nos dados criminais de 2001, utilizando técnicas de Análise Exploratória de Dados Espaciais (AEDE) como o Índice I de Moran e mapas de *clusters* LISA. Seus resultados revelaram uma significativa autocorrelação espacial para diversos tipos de crimes, indicando a tendência de municípios com níveis de criminalidade semelhantes estarem geograficamente próximos, reforçando a necessidade de considerar a distribuição espacial. Essa constatação de Almeida ecoa em diversos outros estudos que aplicaram a AEDE para identificar padrões geográficos da criminalidade em diferentes contextos. Guimarães & Becker (2021), por exemplo, demonstraram a presença de autocorrelação espacial no Rio Grande do Sul, enquanto Andrade (2022) identificou *clusters* de alta criminalidade no Paraná. No próprio estado de São Paulo, Davoglio (2019) também utilizou a AEDE para analisar roubos e furtos de 2016, identificando padrões de concentração e associações com fatores socioeconômicos. Um exemplo de seu trabalho, está na Figura 8. A presente dissertação se alinha com essa robusta abordagem metodológica, buscando aplicar a AEDE para investigar os padrões criminais nos municípios paulistas, utilizando dados mais recentes e abrangentes do período de 2017 a 2024, e assim contribuir para a compreensão da dinâmica espacial da criminalidade no estado.

Figura 8 – Análise bivariada entre roubos e densidade demográfica para o ano de 2016 no estado de São Paulo



Fonte: (Davoglio, 2019)

Outra linha de investigação para a compreensão da dinâmica criminal reside na análise da ligação entre a criminalidade e os fatores socioeconômicos, considerando a sua distribuição espacial. Francia (2024), em sua monografia, aplicou modelos de econometria espacial para examinar a relação entre o mercado de drogas ilícitas e crimes violentos no estado de São Paulo, demonstrando a existência de dependência espacial entre os municípios e a influência de variáveis socioeconômicas como a escolaridade. Essa abordagem realça como a criminalidade em uma localidade pode ser influenciada por seus vizinhos e pelas condições estruturais dos municípios. Em consonância com essa perspectiva, Araújo (2018) desenvolveu um modelo espaço-temporal para analisar furtos e roubos na cidade de Fortaleza, buscando separar os efeitos de fatores comuns (globais) das interações locais (efeito vizinhança). Seus resultados evidenciam a influência de fatores macroeconômicos e políticas públicas, bem como a importância da dinâmica espacial na distribuição dos crimes. No mesmo sentido, Gaulez & Maciel (2016) investigaram os determinantes socioeconômicos da criminalidade em São Paulo com uma análise espacial em corte transversal, encontrando associações positivas entre variáveis como desigualdade de renda, urbanização, densidade populacional e as taxas de criminalidade, além de confirmarem a autocorrelação espacial. Em conjunto, esses estudos sublinham a necessidade de ir além da análise isolada de fatores, incorporando a complexidade das interações espaciais e a influência do contexto socioeconômico para uma compreensão mais aprofundada dos padrões de criminalidade. Essa perspectiva de análise integrada, que considera tanto a dimensão espacial quanto os fatores estruturais, informa a abordagem metodológica da presente dissertação na investigação da criminalidade nos municípios paulistas.

Complementando as análises empíricas, o artigo de Nery & Adorno (2022) oferece uma perspectiva valiosa ao apresentar uma análise retrospectiva e crítica da produção científica sobre criminalidade e violência em São Paulo. Os autores traçam a evolução das abordagens teóricas e metodológicas ao longo das décadas, desde estudos baseados em dados agregados e métodos qualitativos até a crescente adoção de ferramentas quantitativas, estatísticas espaciais e análises de redes sociais. Ao destacar os avanços no acesso a dados públicos e a importância da articulação entre indicadores sociais e criminais, o estudo também aponta limitações persistentes e defende o uso de técnicas de ciência de dados e análises espaço-temporais como caminhos promissores para aprofundar a compreensão da complexa dinâmica criminal nas grandes cidades brasileiras. Essa visão panorâmica e a defesa de abordagens analíticas integradas, com foco nas dimensões espacial e temporal, reforçam a relevância da metodologia exploratória proposta na presente dissertação para o estudo da criminalidade nos municípios do estado de São Paulo, buscando avançar na compreensão dessa complexa dinâmica.

Em suma, a literatura revisada demonstra consistentemente que a criminalidade urbana não é um fenômeno aleatório, mas sim concentrado espacial e temporalmente, sendo influenciado por fatores socioeconômicos e pela dinâmica entre áreas vizinhas. A integração de métodos da ciência de dados, incluindo visualização, análise espacial, estudos revisados focados no contexto paulista, aprendizado de máquina, emerge como uma abordagem promissora para a compreensão

da complexa dinâmica criminal, o que motiva e fundamenta a metodologia exploratória proposta nesta dissertação para o estado de São Paulo.

2.4 CONSIDERAÇÕES FINAIS SOBRE A REVISÃO BIBLIOGRÁFICA

A revisão da literatura evidencia o papel cada vez mais central da ciência de dados no avanço da compreensão dos complexos padrões criminais, dada a sua capacidade de processar grandes volumes de informação e identificar relações sutis. A aplicação de técnicas como a Análise Exploratória de Dados, aliada à crescente incorporação das dimensões espacial e socioeconômica, demonstra um potencial significativo para a formulação de políticas públicas de segurança mais eficazes e fundamentadas em evidências concretas.

Apesar dos notáveis avanços metodológicos observados na literatura, ainda persistem lacunas importantes a serem exploradas, particularmente no contexto brasileiro. Uma dessas lacunas reside na necessidade de uma aplicação mais integrada de diferentes abordagens analíticas em escala municipal, que possa capturar a heterogeneidade e as nuances locais da criminalidade. Este trabalho busca contribuir para o preenchimento dessa lacuna ao aplicar técnicas de análise exploratória sobre um conjunto de dados criminais abrangente dos municípios do estado de São Paulo, compreendendo o período recente de 2017 a 2024, o que permitirá uma análise das tendências criminais mais atuais. A próxima seção detalhará os dados que possibilitarão operacionalizar as discussões teóricas apresentadas, descrevendo suas fontes, as variáveis selecionadas e o processo de tratamento adotado para garantir a qualidade das análises.

3 MATERIAIS E MÉTODOS

Dando prosseguimento à contextualização do problema e à fundamentação teórica apresentadas nos capítulos anteriores, este capítulo detalha a origem, o conteúdo e os critérios de seleção dos dados que serviram de base para a presente investigação sobre a dinâmica da criminalidade nos 645 municípios do estado de São Paulo, no período de 2017 a 2024. Além dos dados, são apresentadas as etapas metodológicas adotadas para a realização das análises dos dados. Para uma compreensão visual do fluxo metodológico que guiou esta pesquisa, a Figura 9 ilustra o pipeline de análise de dados implementado.

Figura 9 – Etapas metodológicas da pesquisa

Fonte: Elaborada pela autora

Vale ressaltar que a qualidade e a relevância dos dados coletados foram cruciais para a análise exploratória da pesquisa, visando identificar padrões espaço-temporais e associações com fatores socioeconômicos. Para tanto, o capítulo se organiza em cinco seções. Inicialmente, serão apresentados os dados criminais, incluindo sua fonte e as variáveis selecionadas; em seguida, serão descritos os dados socioeconômicos utilizados para contextualizar a criminalidade. A terceira seção abordará os dados geográficos, essenciais para a análise espacial. Em seguida, será detalhado o pré-processamento aplicado aos dados para garantir sua adequação às análises subsequentes. Por fim, são apresentas as ferramentas e plataformas utilizadas para obter todos os resultados.

3.1 DADOS DE CRIMINALIDADE

Os dados referentes às ocorrências criminais para os 645 municípios paulistas foram extraídos do portal oficial da Secretaria da Segurança Pública de São Paulo (SSP-SP, 2024). Este portal disponibiliza, de forma aberta e sistematizada, informações mensais sobre diversos tipos de crimes registrados em cada município paulista, oferecendo uma granularidade temporal e espacial essencial para a análise proposta.

A base de dados contempla 9 categorias de crimes disponibilizadas pela SSP-SP (2024):

- Homicídio doloso;

- Homicídio culposo;
- Tentativa de homicídio;
- Lesão corporal seguida de morte;
- Lesão corporal;
- Latrocínio;
- Estupro;
- Roubo;
- Furto.

Para cada município, os registros estão organizados por mês e ano, permitindo a construção de séries temporais por tipo de crime e por localidade.

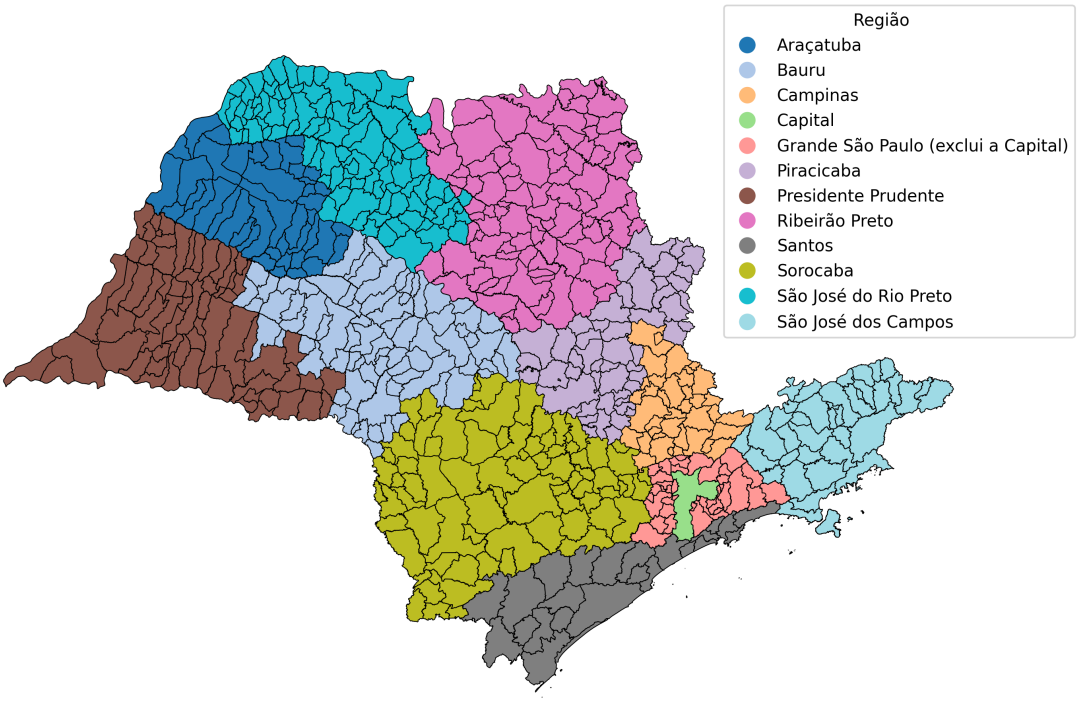
Embora a SSP-SP (2024) disponibilize registros históricos de ocorrências criminais desde 2001, a presente pesquisa delimitou o período de análise entre janeiro de 2017 e dezembro de 2024. Essa escolha temporal se deve principalmente à evolução na forma como a SSP-SP categoriza e reporta alguns tipos de crime ao longo dos anos, impactando a comparabilidade das séries históricas. Especificamente, o ano de 2016 não apresenta o detalhamento das ocorrências de "estupro" e "roubo outros", da mesma forma que o período de 2017 a 2024. Por outro lado, embora o ano de 2015 possuísse as categorias "estupro" e "roubo outros", a categoria "estupro de vulnerável", que se tornou consistente a partir de 2016, não estava presente. Diante dessas mudanças na granularidade e na nomenclatura das categorias de crime, optou-se por restringir a análise ao período de 2017 a 2024, garantindo a uniformidade e a comparabilidade das categorias criminais ao longo do tempo e, consequentemente, a qualidade e a confiabilidade das análises das tendências criminais recentes.

O número de ocorrência de homicídio culposo corresponde a todas as formas de homicídio culposo notificada pela SSP-SP (2024), assim como a categoria de lesão corporal engloba tanto a forma culposa quanto a dolosa, categoria estupro inclui o estupro de vulnerável, roubo se refere a todos os tipo de roubo (banco, veículo, carga e outros) e furto soma furto de veículos e outros. No total, a SSP-SP (2024) disponibiliza 18 tipos de crimes, as análises com todos seriam mais completas, mas seriam extensas e repetitivas detalhar uma por uma, assim optamos por juntar alguns semelhantes.

A base também inclui informações auxiliares, como o código do município (ID) e a região administrativa a que pertence, viabilizando análises agregadas por região ou *clusters* de municípios e facilitando a exploração de padrões espaciais e regionais da criminalidade. Os municípios foram divididos em 12 regiões, pela SSP-SP (2024), sendo elas: Araçatuba, Bauru, Campinas, Capital, Grande São Paulo (exclui a Capital), Piracicaba, Presidente Prudente, Ribeirão Preto, Santos, São José do Rio Preto, São José dos Campos e Sorocaba. A cidade de São Paulo possui

dados discrepantes das demais cidades, comparados até a regiões, sendo assim, foi separada como uma região. A Figura 10 mostra as regiões no mapa do estado de São Paulo.

Figura 10 – Mapa dos municípios de São Paulo por região



Fonte: Elaborada pela autora

3.2 DADOS SOCIOECONÔMICOS

Para complementar a análise dos dados criminais e possibilitar estudos correlacionais, foram coletadas variáveis socioeconômicas dos municípios paulistas. Essas variáveis foram obtidas por meio do IBGE (2023).

As principais variáveis utilizadas incluem:

- População estimada (por ano);
- Área territorial do município (km²);
- Densidade demográfica (habitantes/km²);
- Produto Interno Bruto (PIB) municipal;
- Índice de Desenvolvimento Humano Municipal (IDHM).

O IDHM é composto pela expectativa de vida, educação e renda da população de cada município. Esses valores não são calculados anualmente para todos os municípios brasileiros, uma vez que tais dados são divulgados com base nos Censos Demográficos do IBGE, realizados a cada dez anos. Trabalharemos com o último dado calculado, de 2022.

O PIB municipal é disponibilizado anualmente pelo IBGE, porém com uma defasagem, o de 2021, por exemplo, foi divulgado em dezembro de 2023. Assim, este trabalho considera o PIB de 2017 a 2021. É importante pontuar que o PIB é a soma de todos os bens e serviços finais produzidos dentro de um município ao longo de um ano, e representa o tamanho da economia local, sendo uma das principais medidas do desempenho econômico de uma cidade, região, estado ou até mesmo de um país.

Informações como tamanho da população também dependem dos Censos Demográficos do IBGE, que são realizados a cada dez anos, porém, o IBGE, com uso de métodos estatísticos faz o cálculo da estimativa populacional anualmente. Nesse trabalho, foi adotada a população estimada de 2017 a 2024. De posse da população estimada e da área de cada cidade, foi calculada a densidade demográfica.

Essas variáveis fornecem um panorama geral das condições estruturais e econômicas de cada município, sendo úteis na análise de possíveis fatores associados à criminalidade. Além disso, a combinação entre dados criminais e socioeconômicos permite o cálculo de indicadores derivados, como taxas de criminalidade por 100 mil habitantes, que são fundamentais para uma comparação equitativa entre municípios de diferentes portes populacionais.

3.3 DADOS GEOGRÁFICOS

Para possibilitar a análise da dimensão espacial da criminalidade, foram incorporadas as coordenadas geográficas (latitude e longitude) de cada município do estado de São Paulo, informações essenciais para o georreferenciamento dos dados criminais e socioeconômicos. Adicionalmente, utilizou-se a Malha Municipal Digital (MMD) do Brasil, disponibilizada pelo IBGE (2024). Esta malha define precisamente os limites territoriais de cada um dos 645 municípios paulistas, e é fundamental para a realização de análises espaciais avançadas, como o cálculo de matrizes de pesos espaciais baseadas em contiguidade ou distância, a identificação de vizinhança entre municípios e a aplicação de técnicas de autocorrelação espacial, para verificar se há padrões de aglomeração ou dispersão da criminalidade.

As coordenadas geográficas e a MMD viabilizam a construção de diversas visualizações georreferenciadas, cruciais para a compreensão da distribuição espacial da criminalidade. Entre elas, destacam-se os mapas de calor, que permitem identificar áreas de maior concentração de ocorrências criminais, os mapas coropléticos, que exibem as taxas de criminalidade por município através de diferentes tonalidades de cor, e os mapas de *clusters* espaciais, que identificam agrupamentos de municípios com características criminais semelhantes, como *hotspots* e *coldspots*. Essas informações geográficas são, portanto, fundamentais para alcançar os objetivos desta pesquisa, permitindo compreender a distribuição geográfica da criminalidade entre os municípios de São Paulo, identificar possíveis padrões regionais e investigar se a proximidade geográfica influencia os níveis de criminalidade.

3.4 PRÉ-PROCESSAMENTO DOS DADOS

Antes da realização das análises, foi necessário efetuar uma etapa de pré-processamento dos dados coletados e reunidos para garantir a consistência, completude e qualidade das análises. Esse processo incluiu as seguintes ações:

- Seleção do período temporal: foram considerados os dados compreendidos entre janeiro de 2017 e dezembro de 2024, devido à ausência de dados de algumas categorias de crimes em períodos anteriores, conforme mencionado na seção de dados criminais.
- Integração das bases de dados: os dados criminais da SSP-SP foram anexados aos dados socioeconômicos e geográficos do IBGE, utilizando o código do município (ID). Essa união enriqueceu a base de dados com variáveis explicativas adicionais, permitindo análises mais abrangentes.
- Tratamento de valores ausentes: inicialmente, a base de dados contemplava o período de 2002 à 2024, porém, foram verificados valores faltantes nos dados anteriores à 2017 em algumas categorias específicas de crime. Em especial, foi identificado que não existiam tais categorias em anos anteriores, assim, foi decidido desconsiderar todos os dados anteriores a janeiro de 2017, para não haver inconsistência nas análises. Não houve ausência de dados para as demais variáveis.
- Cálculo de variáveis derivadas: a partir dos dados originais, foram criadas novas variáveis (artificiais) para facilitar a análise e as comparações, conforme a relação abaixo.

    1. Taxa de criminalidade por 100 mil habitantes (T): calculada para normalizar a incidência criminal e permitir a comparação entre municípios com diferentes tamanhos populacionais. A fórmula utilizada foi:

    $$T_i = \frac{N_i \times 100000}{P},$$

    onde N_i representa o total de ocorrências do tipo de crime i, e P é a população estimada do município.

    2. Índice de Criminalidade (IC): a fim de quantificar quão perigoso um local é em relação a outro, é necessário avaliar quais crimes aconteceram, juntamente com "o peso" desses crimes, considerando ainda o total de habitantes daquele local (Prado, 2019). Para isso, foi computado o IC atribuindo pesos diferenciados a cada tipo de crime. A fórmula utilizada foi:

    $$IC = \frac{\sum_{i=1}^{n} P_i \times T_i}{\sum_{i=1}^{n} P_i},$$

onde n é o número de tipos de crime. P_i é o peso atribuído ao crime i (com valores normalizados de 1 a 5, conforme a Tabela 1), e T_i é a taxa de criminalidade por 100 mil habitantes do crime i.

Essa fórmula foi uma adaptação da fórmula $IC = \frac{\sum_{i=1}^{n} P_i \times N_i}{\sum_{i=1}^{n} P_i}$, apresentada por Prado (2019), considerando as taxas de criminalidade ao invés de número de ocorrências, já que a presente pesquisa abrange municípios com diferentes números populacionais, e na investigação promovida por Prado (2019), foi aplicado para bairros. A atribuição de pesos baseou-se no trabalho "Categorização dos Crimes Violentos no Brasil" (Silva; Ramos, 2024), considerando a pena máxima prevista no Código Penal Brasileiro, a gravidade social e psicológica, e a classificação da violência de cada delito. A Tabela 1 apresenta os pesos, normalizados, de 1 a 5, considerados para cada tipo de crime analisado.

Tabela 1 – Pesos atribuídos aos crimes

Tipo de Crime	**Categoria**	**Peso**
Homicídio doloso	Letal	5
Homicídio culposo	Letal	3.5
Tentativa de homicídio	Letal	4
Lesão corporal seguida de morte	Letal	5
Lesão corporal	Violento	2.5
Latrocínio	Letal e patrimonial	5
Estupro	Sexual	4.5
Roubo	Patrimonial com violência	3
Furto	Patrimonial sem violência	1.5

Fonte: Elaborada pela autora

3. Densidade Demográfica: com o objetivo de obter a concentração populacional, foi calculado a densidade demográfica usando valores de população estimada e áreas disponíveis pelo IBGE, abrangendo todos os anos (2017 a 2024) de todos os municípios paulistas. A fórmula empregada é definida por:

$$D = \frac{P}{A},$$

onde P é a população estimada do município e A a área.

- Conversão de datas: as informações temporais, originalmente em formato de mês textual (Janeiro a Dezembro), foram padronizadas para o formato numérico inteiro (1 a 12), facilitando a manipulação e a análise temporal dos dados.

- Detecção e tratamento de *outliers*: uma das etapas da limpeza dos dados, presente no pré-processamento, é a identificação de valores extremos. Para a visualização da distribuição dos dados, foi usado um *boxplot* para cada variável, encontrando valores acima do limite superior e abaixo do limite inferior, pelo método IQR (valores menores que $Q1 - 1,5IQR$ e maiores que $Q3 + 1,5IQR$, onde IQR é o intervalo interquartil $Q3 - Q1$), confirmando

a presença de *outliers*. Os *outliers* podem distorcer estimativas de parâmetros estatísticos, prejudicar a legibilidade dos gráficos e o tratamento adequado permite uma visualização mais clara e interpretável dos dados.

A pesquisa abrange os 645 municípios do estado de São Paulo. A capital paulista, sozinha, concentra aproximadamente 25% da população estadual - um claro valor extremo que comprova a autenticidade dos dados analisados, incluindo os *outliers* observados. Essa distribuição populacional assimétrica corrobora a veracidade dos dados utilizados no estudo. Sendo assim, em algumas análises específicas, foram mantidos os valores extremos, enquanto em outras, onde a interferência prejudicava os resultados, foram tratados de maneira que suavizasse os dados, não influenciavam negativamente nas análises. Dessa forma, foi observado nos resultados quais dados foram utilizados, isto é, tratados ou não.

Nos dados tratados, foi aplicada a transformação logarítmica com o objetivo de suavizar o impacto de valores extremos. A transformação logarítmica, definida como $log(x + 1)$, reduz a amplitude das variáveis, comprimindo os valores mais altos e aproximando-os da média sem alterar a ordem relativa dos dados. Essa abordagem permite atenuar o efeito desproporcional de municípios com valores excepcionalmente elevados (como a capital São Paulo), mantendo a integridade das informações e melhorando a conformidade com pressupostos estatísticos exigidos por técnicas como análise de componentes principais (PCA). A transformação também contribui para tornar as distribuições mais simétricas e adequadas à análise gráfica (Berg et al., 2021).

Esse conjunto de ações, da etapa de pré-processamento, foi fundamental para assegurar a integridade e a robustez da base de dados utilizada nas etapas subsequentes da pesquisa, possibilitando análises confiáveis e interpretações alinhadas ao contexto socioespacial do estado de São Paulo.

3.5 FERRAMENTAS E PLATAFORMAS COMPUTACIONAIS

A execução desta pesquisa exigiu o uso de um conjunto robusto de ferramentas computacionais que permitiram desde a coleta e tratamento dos dados até a visualização e análises estatísticas e espaciais. A escolha das ferramentas foi guiada pela necessidade de flexibilidade, reprodução, integração entre métodos, além de ampla documentação e comunidade de suporte.

Para o desenvolvimento dos *scripts* e análises, foi utilizada a plataforma *Google Colab*, um ambiente interativo baseado em *notebooks Jupyter*, que permite a execução de código *Python* em nuvem, sem necessidade de instalação local (Google Research, 2024). A plataforma foi essencial para a organização do fluxo de trabalho, integração com o *Google Drive*, compartilhamento dos experimentos e reprodutibilidade dos resultados.

A linguagem de programação *Python* foi escolhida por sua versatilidade e por dispor de uma ampla variedade de bibliotecas voltadas à ciência de dados, estatística e geoprocessamento.

Todas as etapas do projeto foram implementadas nessa linguagem. As bibliotecas utilizadas foram:

- *Pandas*: essencial para a manipulação e organização de dados em estruturas tabulares (*DataFrames*) (Müller; Guido, 2016). Foi empregada para a leitura, filtragem, agregação, fusão e tratamento de valores ausentes e inconsistências nos conjuntos de dados.
- *NumPy*: fundamentou as operações numéricas de alta performance, especialmente para manipulação de *arrays* e cálculos matemáticos.
- *Matplotlib* e *Seaborn*: foram empregadas na construção de gráficos exploratórios, como gráficos de linha, *boxplots* e *violin plots*, matrizes de correlação, mapas de calor, e adaptadas para a criação de mapas coropléticos, fundamentais para a análise descritiva e visualização de padrões (McKinney, 2018).
- *Scipy*: utilizada em análises de correlação (como o coeficiente de correlação de Pearson) entre os indicadores criminais e variáveis socioeconômicas.
- *Statsmodels*: empregada para a decomposição de séries temporais, permitindo a separação de componentes como tendência, sazonalidade e resíduo, facilitando a compreensão dos padrões subjacentes aos dados.
- *Geopandas*: foi fundamental para a manipulação de dados geoespaciais e a integração de informações geográficas com dados tabulares. Permitindo a leitura de arquivos *shapefile* e a realização de operações espaciais.
- *PySAL (Python Spatial Analysis Library), em especial o módulo ESDA (Exploratory Spatial Data Analysis)*: utilizadas para a realização de análises de autocorrelação, identificar *clusters* e *outliers* espaciais. Especificamente, foram aplicados para o Índice de Moran e LISA (Rey; Anselin, 2007).
- *Scikit-learn*: foi aplicada na redução de dimensionalidade por meio da Análise de Componentes Principais (PCA) e na identificação de padrões por meio do algoritmo de agrupamento *K-means* (Müller; Guido, 2016).

A combinação dessas ferramentas e bibliotecas proporcionou um ambiente completo e eficiente para a condução das análises propostas, desde a fase exploratória até a obtenção de resultados significativos para a dissertação.

Este capítulo forneceu uma descrição detalhada do conjunto de dados utilizado para cumprir com os objetivos desta dissertação, integrando informações criminais, socioeconômicas e geográficas dos municípios de São Paulo, juntamente com as etapas de pré-processamento que asseguraram sua adequação para análise. De posse da base de dados construída e devidamente tratada, o próximo capítulo se concentrará na apresentação dos resultados obtidos

com a aplicação dos métodos de análise exploratória, das técnicas estatísticas e de aprendizado de máquina, para identificar padrões espaço-temporais relevantes, avaliar a correlação entre indicadores socioeconômicos e a criminalidade, e propor visualizações informativas, conforme delineado nos objetivos desta pesquisa.

4 RESULTADOS E DISCUSSÃO

Com a base de dados detalhada e pré-processada, o presente capítulo dedica-se à apresentação dos resultados e, quando pertinente, dos métodos de análise exploratória aplicados para investigar os padrões da criminalidade nos municípios do estado de São Paulo entre os anos 2017 e 2024.

4.1 ANÁLISE DESCRITIVA INICIAL

Antes de aplicar métodos mais sofisticados, foi realizada uma inspeção descritiva dos dados com o objetivo de compreender a distribuição das variáveis e identificar possíveis valores extremos, assimetrias e padrões gerais.

A Tabela 2 apresenta as estatísticas descritivas de todas as variáveis, com valores de média, desvio padrão, quartis, valor mínimo e máximo. Esta tabela permite uma visão quantitativa mais geral da centralidade e dispersão dos dados, facilitando a interpretação inicial da criminalidade como um todo. Todos os dados considerados foram anuais.

Tabela 2 – Descrição dos dados

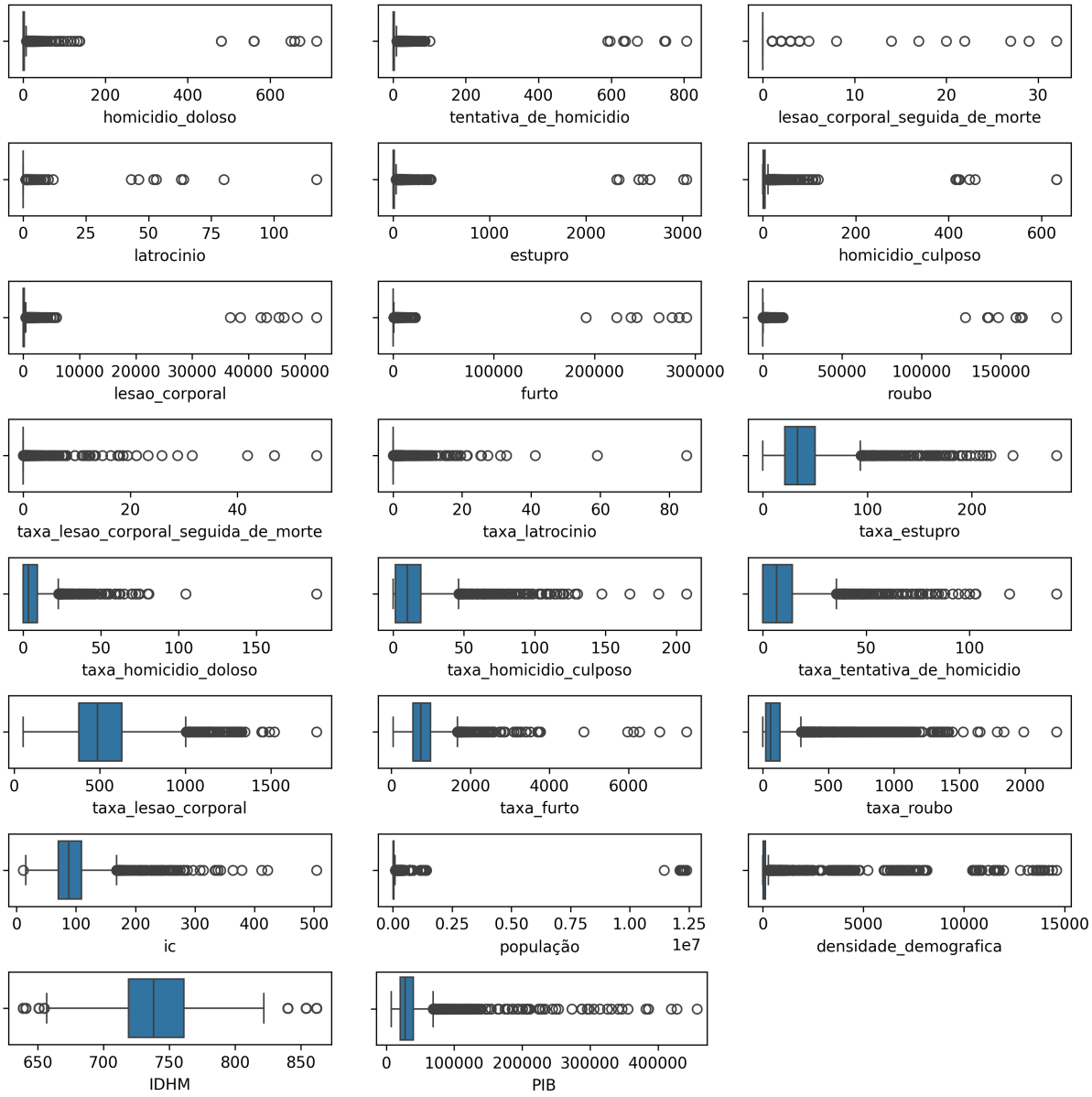
	média	std	min	25%	50%	75%	max
homicidio_doloso	4.390	25.207	0.000	0.000	1.000	3.000	713.000
tentativa_de_homicidio	5.395	28.267	0.000	0.000	1.000	4.000	807.000
lesao_corporal_seguida_de_morte	0.144	0.977	0.000	0.000	0.000	0.000	32.000
latrocinio	0.320	2.791	0.000	0.000	0.000	0.000	117.00
estupro	19.482	108.599	0.000	2.000	5.000	14.000	3041.000
homicidio_culposo	5.610	21.177	0.000	1.000	2.000	5.000	632.000
lesao_corporal	317.264	1812.319	1.000	28.000	67.000	207.250	52074.000
roubo	441.812	6139.475	0.000	1.000	7.000	42.000	185040.000
furto	935.052	10058.653	1.000	36.000	97.000	344.000	291344.000
taxa_lesao_corporal_seguida_de_morte	0.225	1.787	0.000	0.000	0.000	0.000	54.795
taxa_latrocinio	0.410	2.355	0.000	0.000	0.000	0.000	84.962
taxa_estupro	38.728	28.801	0.000	21.121	33.301	50.002	280.899
taxa_homicidio_doloso	6.360	9.567	0.000	0.000	3.350	9.103	188.857
taxa_homicidio_culposo	14.475	17.050	0.000	1.823	10.241	19.618	207.182
taxa_tentativa_de_homicidio	10.026	12.776	0.000	0.000	6.639	14.252	142.045
taxa_lesao_corporal	518.485	196.586	53.022	379.363	488.362	628.931	1767.442
taxa_furto	804.826	426.912	38.948	541.138	744.417	991.626	7470.568
taxa_roubo	121.507	197.834	0.000	21.646	58.525	129.487	2238.979
ic	93.176	36.121	11.089	70.335	87.534	109.420	504.411
população	70740.767	490396.028	836.000	5636.75	13754.500	41284.000	12396372.000
densidade_demografica	338.458	1323.821	3.512	21.349	41.129	123.405	14593.290
IDHM	739.527	32.455	639.000	719.000	738.000	761.000	862.000
PIB	36198.596	34477.243	7572.990	20397.700	27877.780	39783.210	457517.700

Fonte: Elaborada pela autora

A partir dos dados tabulados, é possível observar elevados valores de desvio padrão na maioria das variáveis e uma discrepância considerável ao comparar os valores de 75% com os valores máximos, indicando uma distribuição não normal.

Os *boxplots* apresentados na Figura 11 ilustram a distribuição das variáveis criminais absolutas e normalizadas (taxas), assim como as variáveis socioeconômicas. Os dados usados para a construção do *boxplot* foram anuais.

Figura 11 – *Boxplot de todas as variáveis*



Fonte: Elaborada pela autora

A presença de caudas longas e pontos isolados confirma a existência de fortes assimetrias e *outliers*, especialmente nas variáveis de criminalidade absoluta, como furto, roubo e lesão corporal, traduzindo os valores observados na Tabela 2.

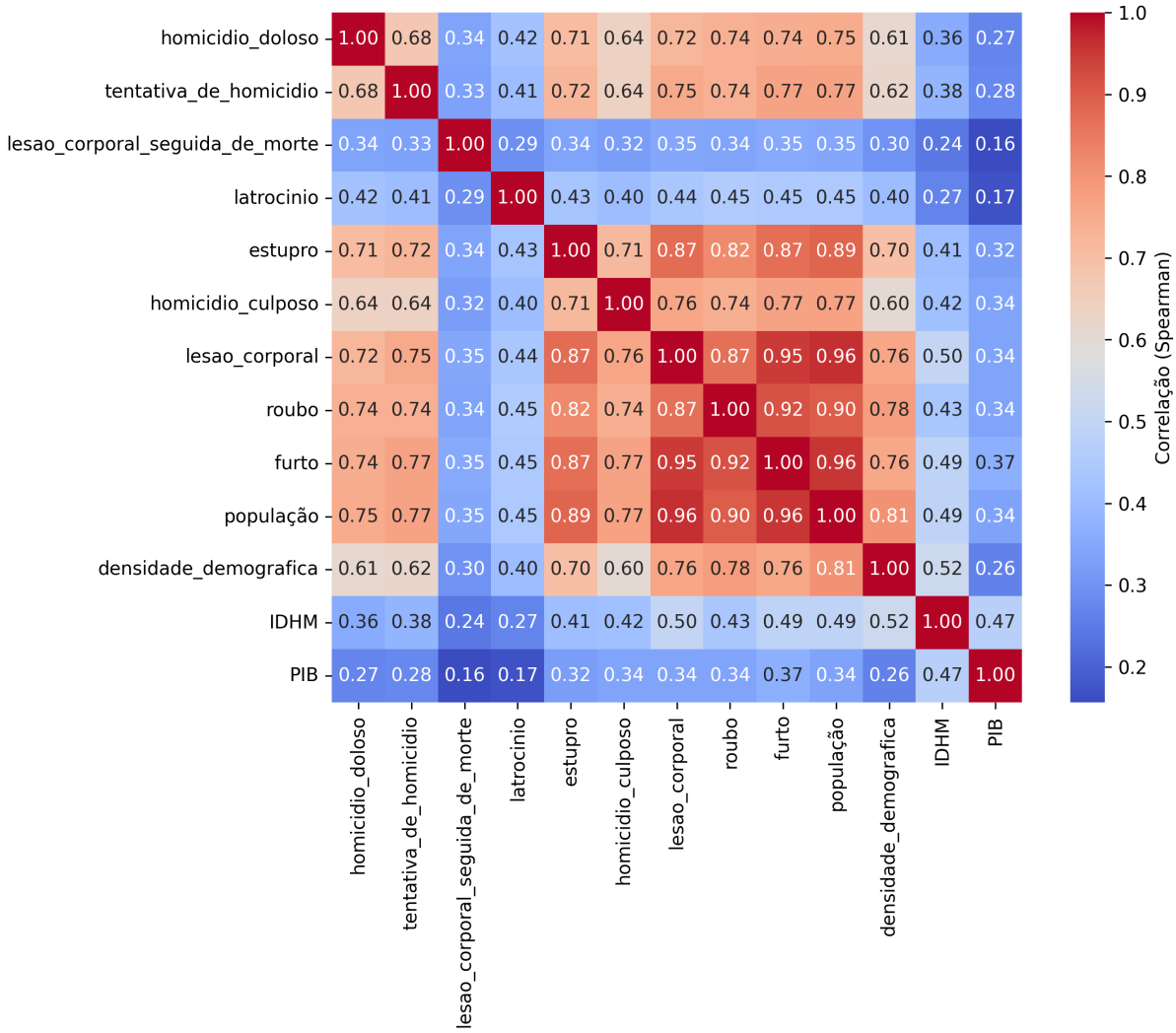
Esses resultados vão ajudar na escolha de alguns dos métodos de análises mais adequados para cada finalidade, como a técnica de correlação.

4.2 ANÁLISES DE CORRELAÇÕES

Para investigar a relação entre os diferentes tipos de crimes e variáveis socioeconômicas, foi calculado o coeficiente de correlação de Spearman. Este método, não paramétrico, foi escolhido por sua robustez frente a *outliers* e por não assumir linearidade entre as variáveis, tornando-o

adequado para o perfil dos dados em análise, que apresenta grande variabilidade e assimetria. A Figura 12 apresenta o mapa de calor de correlação para visualizar os resultados e melhor realizar as análises pertinentes.

Figura 12 – Matriz de correlação de Spearman



Fonte: Elaborada pela autora

A partir do mapa de correlação construído, é possível visualizar as interações entre as variáveis socioeconômicas e números de ocorrências dos 9 tipos de crimes. Os valores variando de 0.16 a 1 são os coeficientes de Spearman para cada par de variáveis, sendo possível observar as interações pelas diferentes tonalidades de cores. Em geral o coeficiente de Spearman varia de -1 a 1, passando de correlação negativa (-1), correlação nula (0) e correlação positiva (1). Nesse caso, é possível observar que não há correlações negativas, e as tonalidades mais próximas do vermelho indicam maiores correlações positivas, o azul mais intenso indica correlação baixa. Aqui destaca-se a alta correlação entre furto, roubo e lesão corporal, com coeficientes em torno de 0.90, indicando que essas ocorrências tendem a crescer em conjunto nos municípios paulistas. Também foi observada correlação significativa entre crimes com maiores números

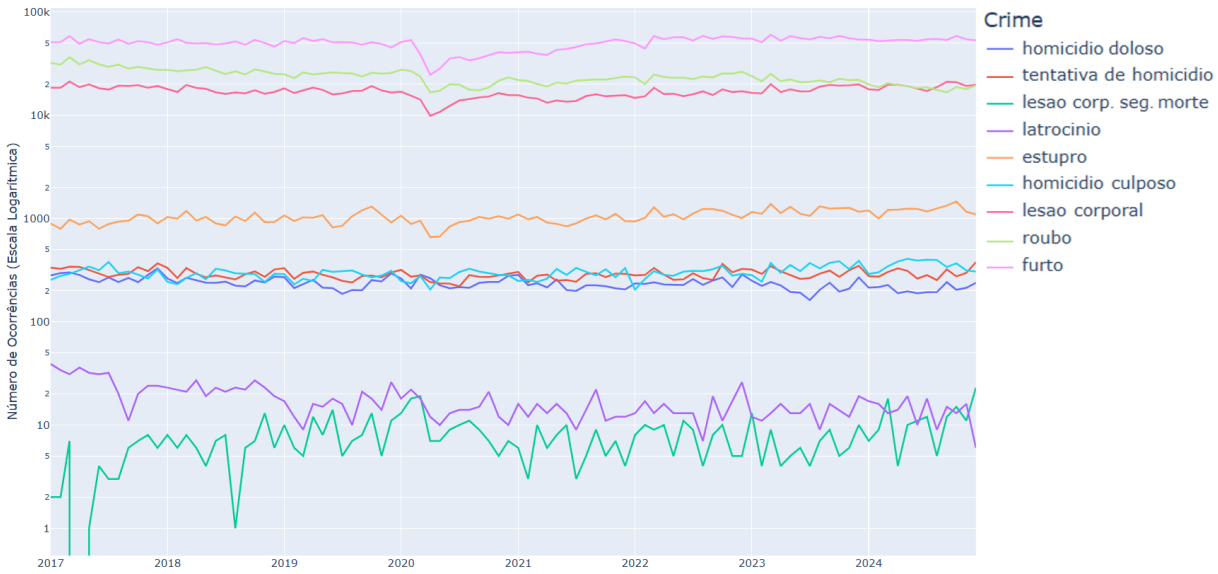
de ocorrências e variáveis demográficas, evidenciado pelos pares população-furto (0.96) e densidade demográfica-roubo (0.78), mostrando o papel da concentração urbana na criminalidade. Crimes como latrocínio, lesão corporal seguida de morte, que são crimes com baixo número de ocorrências (se comparado a roubo e furto) possuem baixa correlação com outras variáveis. Por outro lado, variáveis como o IDHM e PIB apresentaram correlações mais moderadas com os crimes. É importante ressaltar que nem sempre uma alta correlação indica causalidade, ou seja, não é porque o município é muito populoso que vai ocorrer mais furto, por exemplo, podem ter outros fatores que influenciam.

4.3 ANÁLISE DE SÉRIES TEMPORAIS

No tratamento preliminar dos dados, procedemos com o cálculo das taxas criminais por 100 mil habitantes em cada município, estabelecendo assim parâmetros proporcionais que permitem comparações mais equitativas entre localidades com diferentes tamanhos populacionais. Essa abordagem baseada em taxas demograficamente ajustadas fornece bases metodológicas mais sólidas para a análise espacial da criminalidade. Para as análises temporais sem recorte geográfico específico, optamos por utilizar os números absolutos de ocorrências em vez das taxas normalizadas.

Com o objetivo de analisar a evolução temporal dos diferentes crimes no estado de São Paulo, foi plotado um gráfico de linhas (Figura 13). Ele apresenta a série histórica mensal das nove categorias criminais entre 2017 e 2024. Os dados, agregados em nível estadual, foram plotados em escala logarítmica no eixo vertical, solução que permite a visualização integrada de crimes com magnitudes drasticamente diferentes, como os frequentes furtos e os raros latrocínios, sem comprometer a interpretação da série.

Figura 13 – Evolução temporal dos crimes



Fonte: Elaborada pela autora

A análise da série temporal revela padrões distintos entre os tipos de crime. Furto e roubo são as categorias com maior número de registros, mantendo-se consistentemente nas faixas superiores do gráfico ao longo de todo o período analisado. A curva de furtos exibe uma tendência relativamente estável, com variações menores, enquanto roubos demonstram um leve declínio nos anos mais recentes, especialmente após 2020. Crimes com menor frequência, como latrocínio e lesão corporal seguida de morte, aparecem na base do gráfico com flutuações mais expressivas e curvas menos regulares, em parte devido ao baixo volume de registros mensais. Ainda assim, nota-se certa estabilidade após 2021, sem crescimento relevante.

Um aspecto observado é a queda abrupta nas ocorrências de quase todos os crimes em março de 2020, coincidindo com o início da pandemia de COVID-19 e as medidas de isolamento social. Essa redução foi mais acentuada nos crimes contra o patrimônio (furtos e roubos), o que é coerente com a redução da mobilidade urbana nesse período. A recuperação gradual dos números a partir de 2021 sugere o restabelecimento das rotinas sociais.

A maioria dos crimes apresenta aumento especialmente nos meses de março, sugerindo possível influência de fatores sazonais, como datas festivas ou maior circulação de pessoas.

O crime de estupro mostra um pequeno aumento após 2021, mas com menor sazonalidade, enquanto lesões corporais, homicídios culposos, homicídios dolosos e as tentativas de homicídio mantêm uma estabilidade relativa, com pequenas oscilações e sem indícios de crescimento acelerado.

Se comparado com o acumulado de 2017, em 2024 o número total de ocorrências de alguns crimes diminuíram, como latrocínio que teve uma queda de 50%, roubo, 39,17% e homicídio doloso, 23,5%. Outros tiveram um aumento considerado, como estupro 31,5%, homicídio culposo, 16,35% e lesão corporal seguida de morte que aumentou quase 180%.

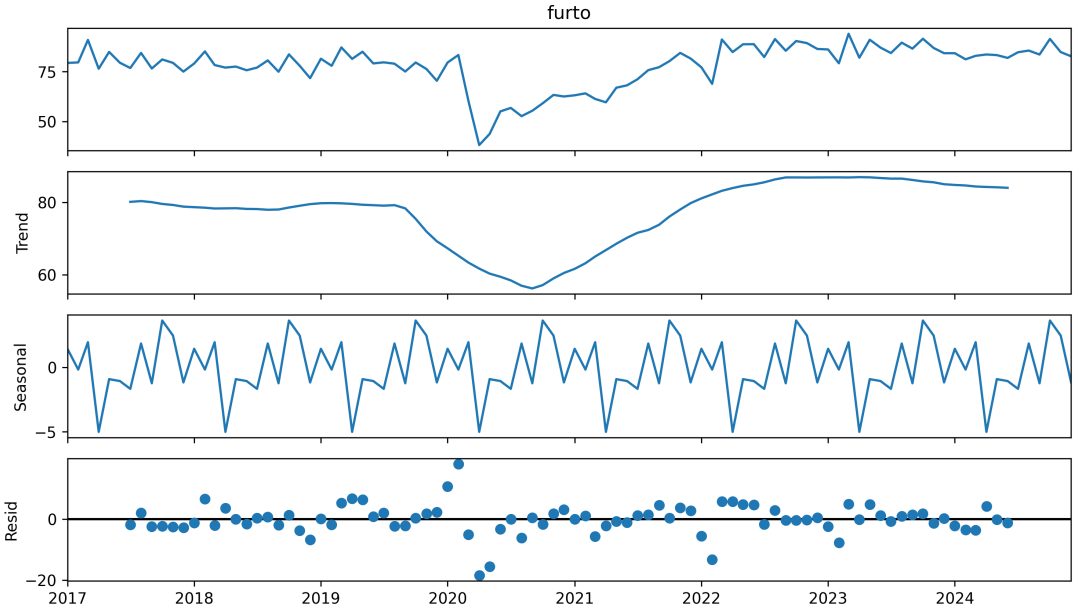
Por fim, o gráfico também revela a heterogeneidade nas magnitudes dos diferentes crimes, mesmo após a transformação logarítmica, reforçando a importância de tratamentos estatísticos diferenciados para cada categoria e a relevância da segmentação regional e temporal em análises subsequentes.

A fim de aprofundar a análise temporal dos crimes, foi realizada a decomposição das séries mensais no período de 2017 a 2024, separando-as em três componentes: i) tendência, ii) sazonalidade, e iii) resíduos. Esse procedimento permite isolar padrões de longo prazo e comportamentos cíclicos recorrentes, facilitando a interpretação de eventos atípicos, como a pandemia, e a identificação de sazonalidade nos dados.

A Figura 14 traz a decomposição da série das ocorrências de furtos. O gráfico de decomposição revela padrões distintos entre os componentes analisados. O gráfico superior apresenta a série original, com o número mensal de ocorrências variando entre aproximadamente 55 e 95 registros, demonstrando flutuações consideráveis ao longo do período. O componente tendência (segundo painel) exibe uma variação suave, iniciando em torno de 80 ocorrências, com um declínio progressivo até meados de 2020, quando atinge o valor mínimo de aproximadamente 57 ocorrências. A partir de então, observa-se uma recuperação gradual, retomando os patamares

anteriores até 2023, e mantendo-se estável até 2024. Este comportamento pode refletir o impacto de fatores externos, como efeitos da pandemia de COVID-19. O componente sazonal (terceiro painel) apresenta oscilações regulares em torno de zero, com amplitude variando entre -5 e 5 ocorrências, indicando a existência de um padrão sazonal, porém de impacto moderado em comparação com a tendência geral da série. A constância e regularidade das oscilações sugerem que o crime de furto possui certa repetitividade ao longo do ano, possivelmente relacionado a fatores cíclicos, como datas comemorativas ou períodos de férias. Por fim, os resíduos (painel inferior) apresentam variações de até ± 20 ocorrências, superiores à amplitude da sazonalidade, refletindo a presença de fatores aleatórios ou eventos pontuais que não foram capturados pela tendência ou pela sazonalidade. Esse comportamento indica que, embora existam padrões estruturais na série, o furto é influenciado por eventos inesperados ou não recorrentes.

Figura 14 – Decomposição de série temporal - furto

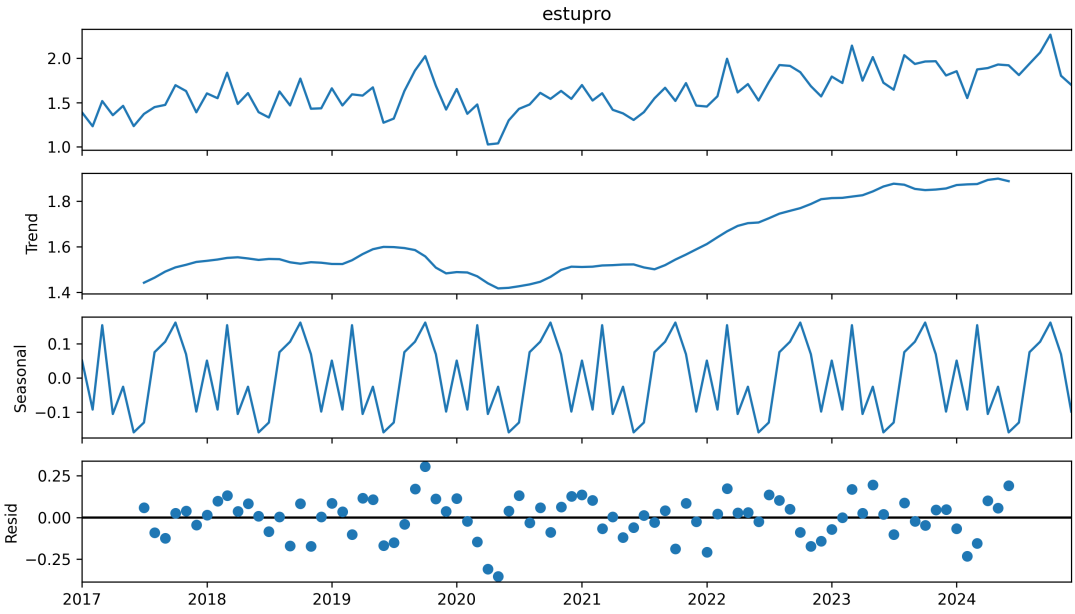


Fonte: Elaborada pela autora

Na decomposição de série temporal de estupro, Figura 15, observa-se um comportamento diferente. No primeiro painel, a série original, nota-se um comportamento oscilatório, com variações ao longo do período analisado (2017 a 2024). A tendência estimada (segundo painel) aponta um crescimento gradual das taxas médias de estupro no estado, principalmente a partir de 2021, indicando um aumento consistente do fenômeno nesse intervalo.

O componente sazonal (terceiro painel) evidencia flutuações regulares, com padrão que se repete ao longo dos anos, porém, com pequena amplitude, o que indica que o efeito sazonal é fraco. Além disso, os resíduos (quarto painel) têm uma variação maior que a sazonalidade, o que significa que há grande variação não explicada, ou seja, o comportamento do crime é mais influenciado por fatores aleatórios ou outros não capturados pelo modelo do que pela sazonalidade.

Figura 15 – Decomposição de série temporal - estupro



Fonte: Elaborada pela autora

De modo geral, a decomposição aponta para uma tendência de crescimento do estupro no estado, com pouco efeito sazonal.

4.4 ANÁLISE ESPACIAL

Visando aprofundar a análise sobre a criminalidade nos municípios paulistas, foi elaborado um ranking com os dez municípios que apresentaram as maiores médias do Índice de Criminalidade (IC) ao longo do período analisado.

O IC consiste em uma métrica composta, calculada a partir das taxas criminais padronizadas por 100 mil habitantes, ponderadas pela gravidade de cada tipologia delitiva.

Essa metodologia possibilita comparações válidas entre municípios de diferentes portes populacionais, identificando aqueles com os mais altos níveis de criminalidade relativa quando considerada a proporção da população local.

Tabela 3 – Municípios com maiores médias de índice de criminalidade

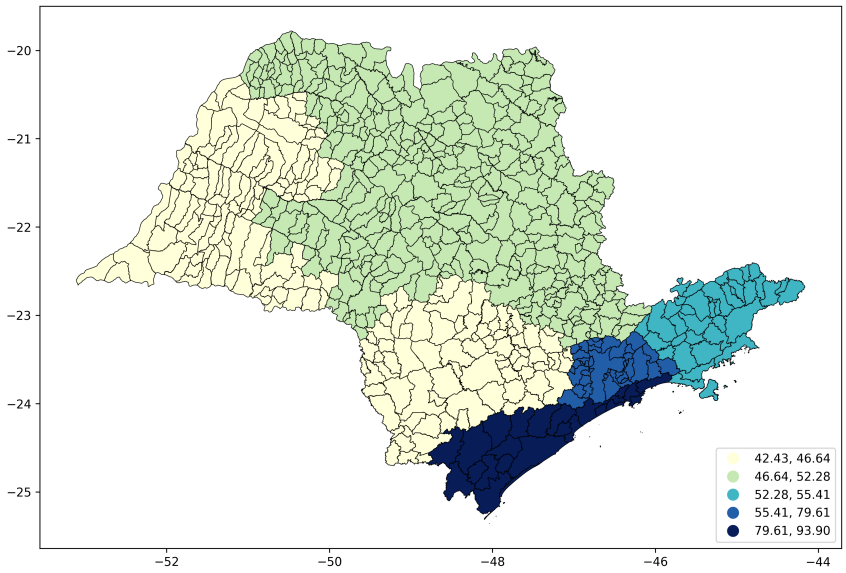
Município	IC
Ilha Comprida	366.648
Mongaguá	305.971
Itanhaém	245.018
Peruíbe	240.359
São Paulo	235.052
Praia Grande	221.199
Diadema	212.621
Miracatu	208.473
Santo André	203.928
Bertioga	193.214

Fonte: Elaborada pela autora

A Tabela 3 destaca os dez municípios com os maiores valores médios de IC. Observa-se um padrão relevante: Ilha Comprida, Mongaguá, Itanhaém, Peruíbe, Praia Grande e Bertioga são municípios da região de Santos e são do litoral paulista. Essa concentração sugere a influência de fatores locais, como sazonalidade turística, características específicas das cidades litorâneas. A capital e cidades vizinhas, como Diadema e Santo André, que pertencem à região da Grande São Paulo, reforçam a tendência de que áreas metropolitanas concentram indicadores elevados de criminalidade.

A partir do cálculo da média de IC por região, em 2024, obtivemos resultados que reforçam a análise conduzida anteriormente: altos índices de criminalidade na região de Santos e Grande São Paulo, e IC menores no interior do estado, como pode ser constatado na Figura 16.

Figura 16 – Média de índice de criminalidade por região

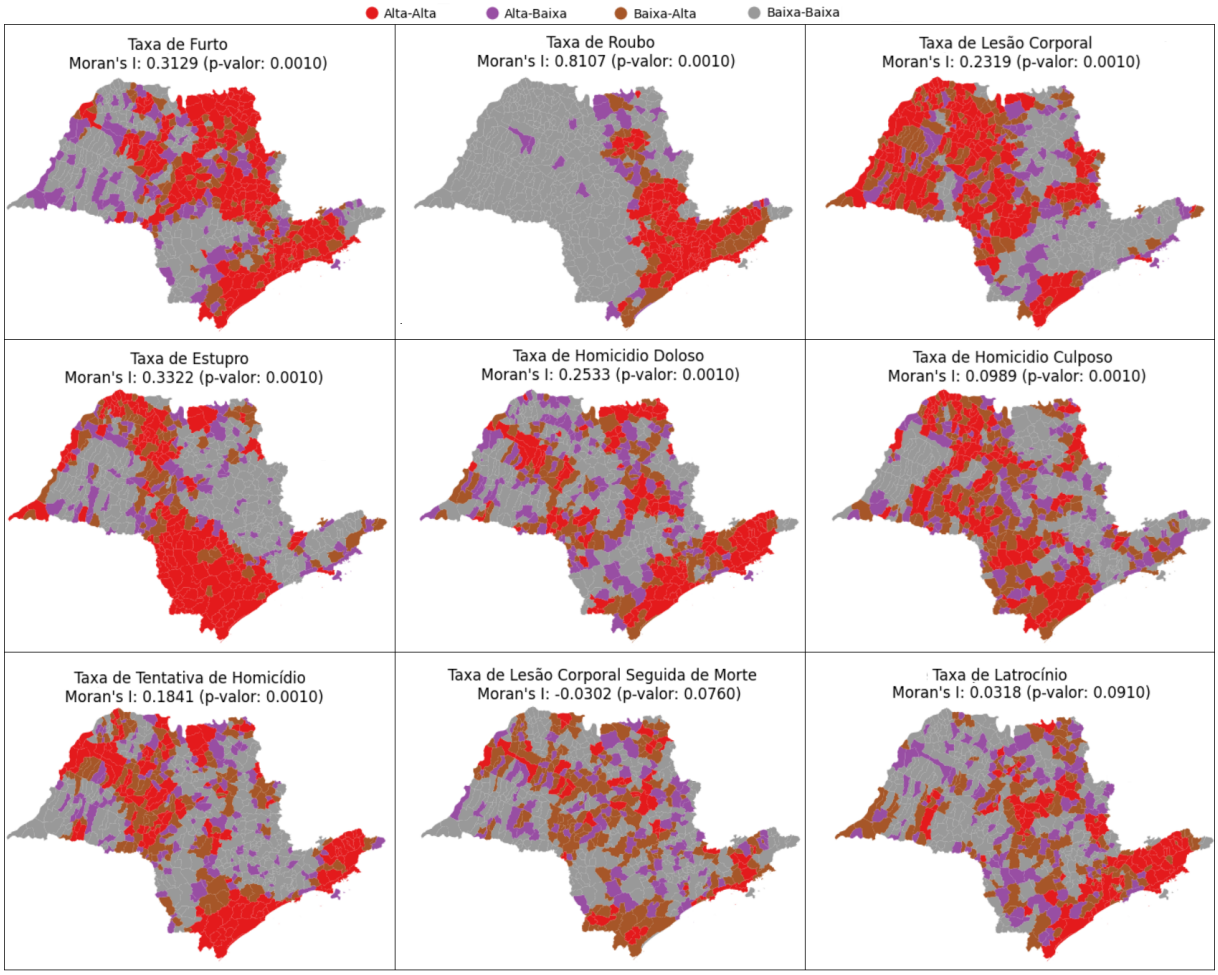


Fonte: Elaborada pela autora

Com o objetivo de identificar onde cada tipo de crime apresenta maior probabilidade de ocorrência, foram realizados agrupamentos espaciais utilizando a técnica LISA – *Local Indicators of Spatial Association* (Almeida, 2007). A Figura 17 apresenta os resultados para as taxas dos nove tipos de crimes investigados no estado de São Paulo, considerando a escala municipal. Para cada crime, foi calculado o Índice I de Moran (Almeida, 2007), que quantifica o grau de autocorrelação espacial, seguido da representação dos *clusters* locais. Esses agrupamentos foram categorizados como: Alta-Alta (municípios com alta taxa cercados por outros com taxas igualmente altas), Baixa-Baixa (municípios com baixa taxa próximos a outros de baixa taxa), Alta-Baixa (municípios com alta taxa cercados por municípios com baixa taxa) e Baixa-Alta (municípios com baixa taxa cercados por municípios com alta taxa). Essa abordagem permite revelar padrões espaciais estatisticamente significativos, destacando regiões de concentração ou dispersão de crimes.

Os resultados obtidos para as nove categorias de crimes analisados revelam padrões distintos de autocorrelação espacial. No referido gráfico, é importante ressaltar que variáveis com p-valor

Figura 17 – LISA: *clusters* espaciais para taxas de 9 tipos de crimes no estado de São Paulo



Fonte: Elaborada pela autora

≤ 0.05 são consideradas significativas e o índice I de Moran varia entre 1 e -1.

Roubo apresentou o maior índice de Moran (I = 0.8107; p-valor = 0.001), indicando forte autocorrelação espacial. O mapa LISA destaca amplas regiões de *clusters* Alta-Alta (vermelho), concentradas sobretudo na Região da Grande São Paulo e entorno, sinalizando que municípios vizinhos compartilham elevadas taxas de roubo.

Os crimes de Furto (I = 0.3129), Estupro (I = 0.3322), Homicídio Doloso (I = 0.2533) e Lesão Corporal (I = 0.2319) também apresentaram autocorrelação espacial positiva significativa (p-valor = 0.001), com destaque para aglomerações de municípios com altas taxas próximas entre si (Alta-Alta). Isso indica que esses crimes tendem a ocorrer em padrões territoriais definidos, e não de forma aleatória no espaço.

Tentativa de Homicídio (I = 0.1841) e Homicídio Culposo (I = 0.0989; p-valor = 0.001) apresentaram autocorrelação espacial positiva, embora mais fraca, sugerindo um padrão espacial menos definido, mas ainda relevante.

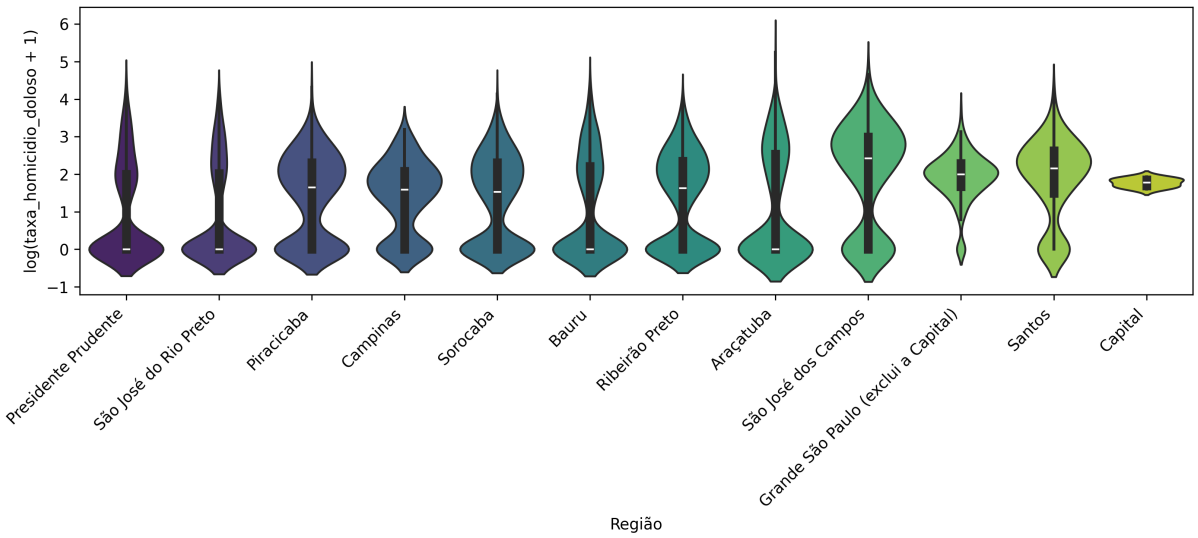
Por outro lado, os crimes de Latrocínio (I = 0.0318; p-valor = 0.091) e Lesão Corporal Seguida de Morte (I = -0.0302; p-valor = 0.076) não apresentaram autocorrelação espacial estatisticamente

significativa. Esses resultados sugerem que essas ocorrências são mais esporádicas ou pontuais, sem formar padrões claros de concentração territorial.

De modo geral, os mapas LISA permitem visualizar que os padrões espaciais variam conforme a tipificação do crime, com alguns apresentando fortes *clusters* de altas taxas — especialmente crimes patrimoniais como furto e roubo — enquanto outros, como latrocínio, mostram distribuições mais aleatórias.

Para analisar a distribuição espacial das taxas de crimes entre as regiões administrativas do estado de São Paulo, foram utilizados gráficos do tipo *Violin Plot*, que combinam a estrutura de um *boxplot* com a curva de densidade estimada. Esses gráficos permitem visualizar não apenas medidas-resumo, como a mediana e os quartis, mas também a forma completa da distribuição dos dados. Para evitar repetições e manter a objetividade da análise, são apresentados apenas dois exemplos: um referente à taxa de homicídio doloso e outro à taxa de roubo. Ambas as variáveis, expressas por 100 mil habitantes, foram transformadas pela função logarítmica $log(x + 1)$, com o objetivo de atenuar a influência de valores extremos e possibilitar uma comparação mais justa entre as regiões. A largura do *Violin Plot* em cada ponto do eixo vertical representa a densidade de dados naquele valor: quanto maior a largura do gráfico, maior a concentração de municípios com aquela taxa específica.

Figura 18 – Distribuição da taxa de homicídio doloso por região



Fonte: Elaborada pela autora

A Figura 18 evidencia a presença de grande assimetria em diversas regiões, indicando que a maioria dos municípios apresenta taxas relativamente baixas, embora com alguns casos muito elevados. A capital paulista, por exemplo, apresenta uma distribuição mais concentrada, com pouca variabilidade e valores consistentemente altos. Isso se justifica por incluir apenas o município de São Paulo, o mais populoso do estado e com taxas médias elevadas de homicídio doloso.

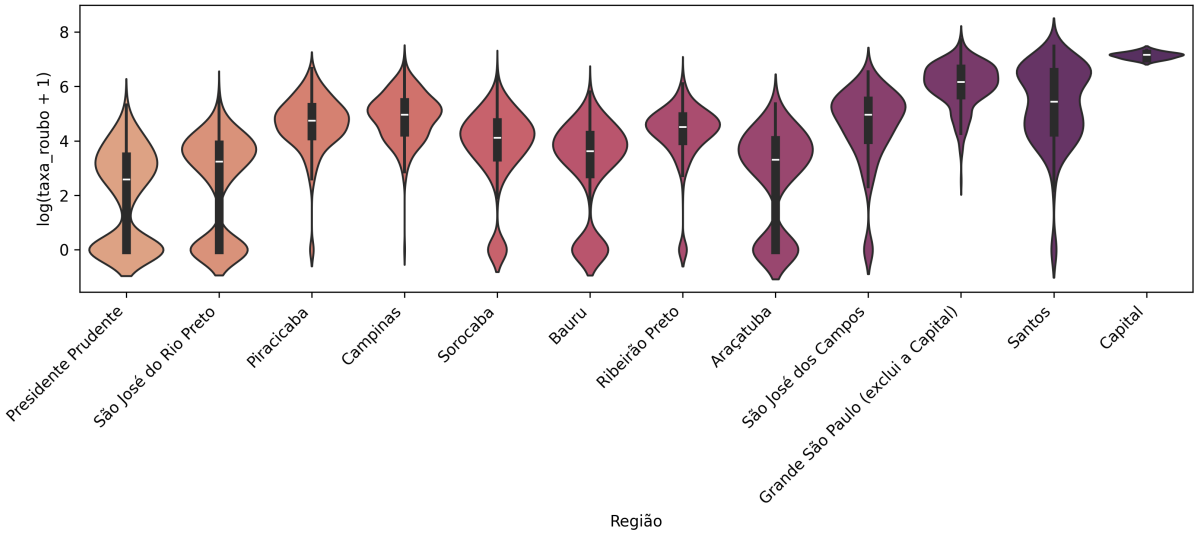
As regiões da Grande São Paulo (exclui a Capital), Santos, São José dos Campos, Piracicaba,

Campinas e Sorocaba também apresentam distribuições com medianas mais altas, indicando maior prevalência de homicídios dolosos. No entanto, como estas regiões reúnem diversos municípios, observa-se maior variabilidade em relação à capital, apresentando alguns municípios com taxas próximas de zero.

Regiões como Presidente Prudente, São José do Rio Preto, Bauru e Araçatuba apresentam distribuições mais concentradas em valores baixos, com muitas cidades registrando taxas próximas a zero, o que significa que mais da metade dos municípios dessas regiões registraram taxas muito baixas ou nulas de homicídio doloso ao longo dos anos analisados. Apesar disso, o formato alargado na parte superior do violino indica que há municípios com taxas mais elevadas, embora sejam casos menos frequentes. Esse contraste entre a mediana baixa e a cauda superior mais larga reflete a desigualdade interna das regiões quanto à ocorrência desse crime.

A maior dispersão em regiões com muitos municípios reforça a heterogeneidade espacial da violência letal no estado.

Figura 19 – Distribuição da taxa de roubo por região



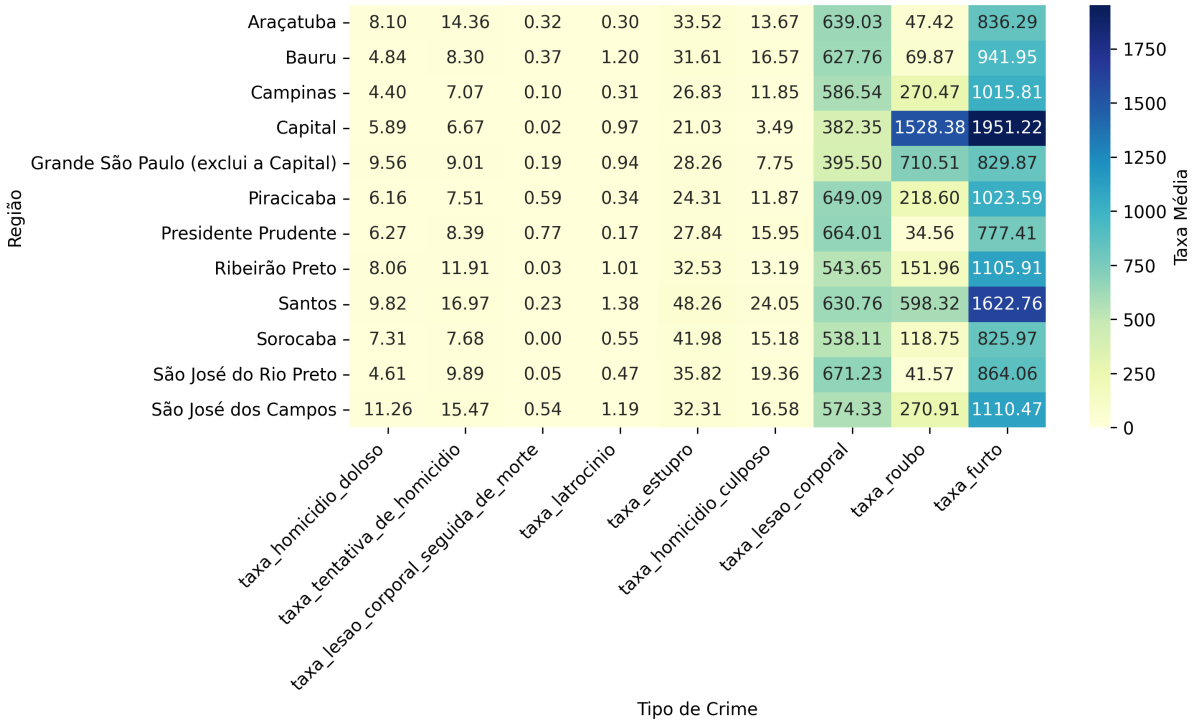
Fonte: Elaborada pela autora

Na Figura 19, de forma geral, observa-se que as taxas de roubo são mais elevadas e variáveis do que as de homicídio doloso. A capital São Paulo, novamente, se destaca com valores altos e baixa dispersão, mantendo a coerência com seu perfil urbano e populacional. Já a Grande São Paulo (exclui a Capital), Santos, São José dos Campos, Piracicaba e Campinas exibem as maiores medianas e valores mais elevados e concentrados, com poucos valores baixos, apresentando uma cauda fina para baixo, revelando uma concentração significativa de municípios com elevados índices de roubo.

Por outro lado, regiões como Presidente Prudente, São José do Rio Preto, Araçatuba, Bauru e Sorocaba apresentam medianas um pouco mais baixas, e um maior volume de municípios com taxas próximas de zero, sugerindo menor incidência desse tipo de crime.

Para compreender a evolução dos padrões criminais ao longo do tempo nas diferentes regiões administrativas do estado de São Paulo, foram construídos mapas de calor representando as médias anuais das taxas dos nove crimes por região. Para fins de comparação, a Figura 20 e Figura 21 mostram os valores referentes aos anos de 2017 e 2024, considerando a média das taxas dos crimes por 100 mil habitantes em cada uma das 12 regiões do estado. O uso de mapas de calor permite identificar visualmente quais regiões concentram os maiores ou menores níveis médios de criminalidade por tipo de crime, além de verificar se esses padrões se modificaram ao longo dos anos.

Figura 20 – Média das taxas de crimes por região - 2017



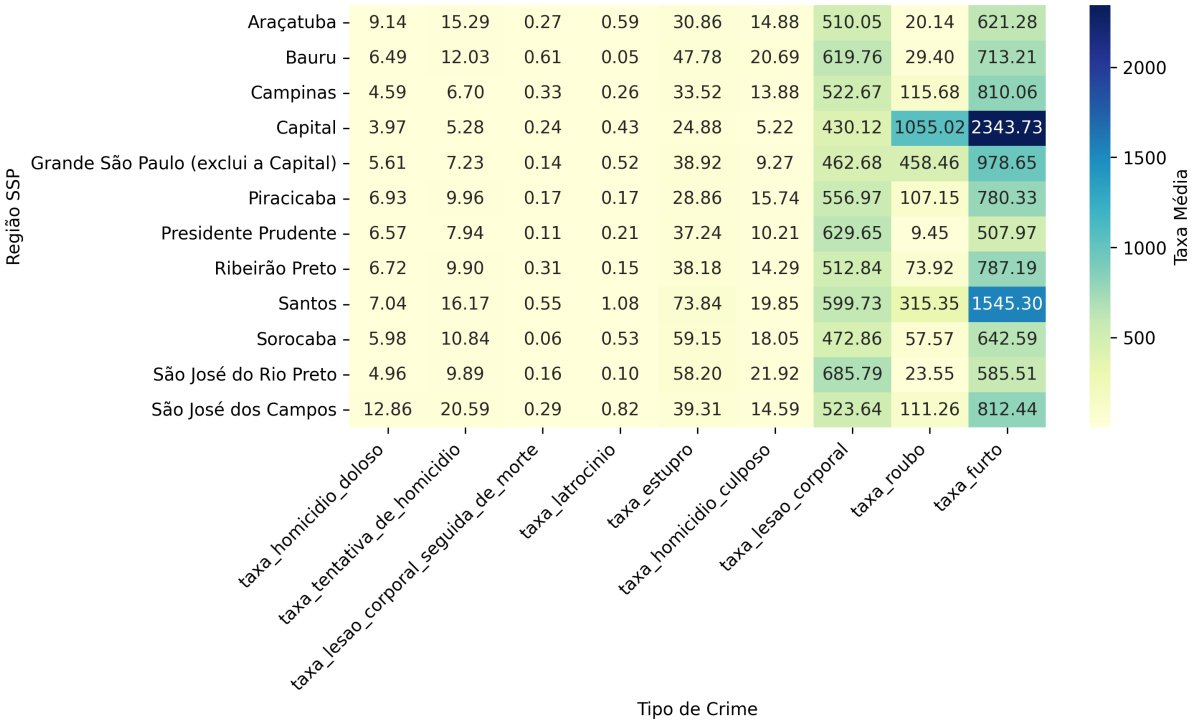
Fonte: Elaborada pela autora

- Furto: em 2017, a Capital apresentou a maior média (1951.22), seguida pela região de Santos (1622.76) e Grande São Paulo (829.87). Em 2024, houve aumento expressivo na Capital, chegando a 2343.73, reforçando a tendência de crescimento. No entanto, nas demais regiões, as médias tiveram uma redução.

- Roubo: em 2017, a Capital registrou a maior média (1528.38), seguida por Santos (598.32) e Grande São Paulo (710.51). Em 2024, a Capital atingiu 1055.02, evidenciando uma queda significativa. As demais regiões também apresentaram redução considerável. Apesar da queda, o roubo continua concentrado principalmente nas áreas metropolitanas.

- Estupro: as taxas de estupro variaram pouco entre os anos analisados. Em 2024, a região de Santos (73.84) manteve-se com as maiores médias, seguidas por Sorocaba (59.15) e São José do Rio Preto (58.20). Houve pequenas oscilações, com uma tendência de aumento.

Figura 21 – Média das taxas de crimes por região - 2024



Fonte: Elaborada pela autora

- Homicídio Doloso: a região de São José dos Campos apresentou as maiores médias nos dois anos (11.26 em 2017 e 12.86 em 2024). A maioria das demais regiões se manteve ou teve um leve aumento nas médias. O crime de homicídio doloso manteve-se em níveis médios estáveis, sem grandes variações.

- Tentativa de Homicídio: São José dos Campos e Santos seguem entre as regiões com maiores taxas médias. As médias se mantiveram relativamente estáveis, com pequenas variações regionais.

- Latrocínio: taxas baixas em todas as regiões, sem grandes variações. Mantém-se um crime de ocorrência rara frente aos demais tipos analisados.

- Lesão Corporal Seguido de Morte: assim como latrocínio, é um crime com baixas taxas e sem muitas variações entre regiões e ao decorrer dos anos.

- Homicídio Culposo: pequenas oscilações, mas sem mudanças de padrão. Santos e São José do Rio Preto mantêm taxas mais elevadas.

- Lesão Corporal: a Capital, região da Grande São Paulo e São José do Rio Preto, tiveram aumento, enquanto as demais regiões apresentaram leve redução. Houve uma tendência geral de leve queda, com regiões do interior como São José do Rio Preto e Presidente Prudente marcando as maiores médias.

De modo geral, observou-se uma redução na maioria dos crimes e na maior parte das regiões administrativas. Destaca-se que os crimes patrimoniais apresentam maior incidência na Grande São Paulo e na Capital, enquanto os crimes contra a vida tendem a apresentar uma distribuição mais homogênea entre as demais regiões do estado.

4.5 ANÁLISE MULTIVARIADA

A fim de compreender os padrões estruturais e regionais dos indicadores criminais no estado de São Paulo, foi aplicada a Análise de Componentes Principais (PCA) com base nas taxas de ocorrência dos nove tipos de crimes por 100 mil habitantes, além de variáveis como IDHM, população e densidade demográfica.

A redução de dimensionalidade promovida pela PCA possibilita visualizar os municípios em um espaço de duas dimensões que conserva a maior parte da variância dos dados. As componentes principais foram interpretadas com base nas cargas fatoriais da Tabela 4, evidenciando quais tipos de crime e fatores estruturais mais influenciam a distribuição dos municípios. Essas cargas são calculadas a partir dos autovetores da matriz de covariância. A seguir, realizou-se uma análise de agrupamento nesse espaço projetado, permitindo identificar perfis distintos de regiões com características criminais e socioeconômicas semelhantes. Os resultados foram representados em gráfico de dispersão, Figura 22, e mapas coropléticos, Figura 23.

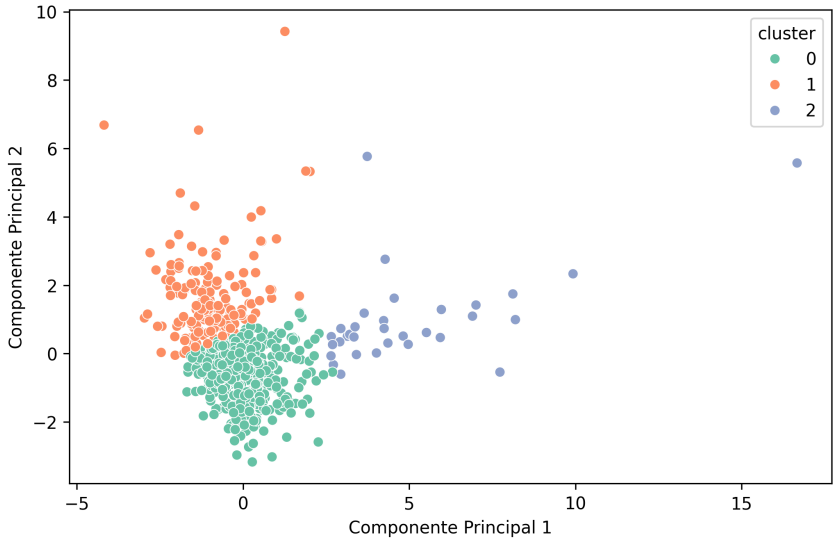
Tabela 4 – Carga dos componentes

Variáveis	PC1	PC2
homicidio_doloso	-0.040	0.431
tentativa_de_homicidio	-0.186	0.471
lesao_corporal_seguida_de_morte	-0.024	0.113
latrocinio	0.033	0.161
estupro	-0.309	0.344
homicidio_culposo	-0.269	0.193
lesao_corporal	-0.130	0.231
roubo	0.498	0.283
furto	0.212	0.477
população	0.346	0.128
densidade_demografica	0.478	0.092
IDHM	0.369	-0.109

Fonte: Elaborada pela autora

A análise por intermédio da técnica PCA revelou que as duas primeiras componentes principais explicam juntas aproximadamente 37,6% da variância total dos dados (PC1: 21,35%; PC2: 16,30%). A primeira componente (PC1) está associada à densidade demográfica, população e taxa de roubo, sugerindo que representa um eixo relacionado à urbanização e exposição a crimes patrimoniais. Já a segunda componente (PC2) se correlaciona com taxas de homicídio doloso, tentativa de homicídio e estupro, apontando para uma dimensão ligada à violência interpessoal.

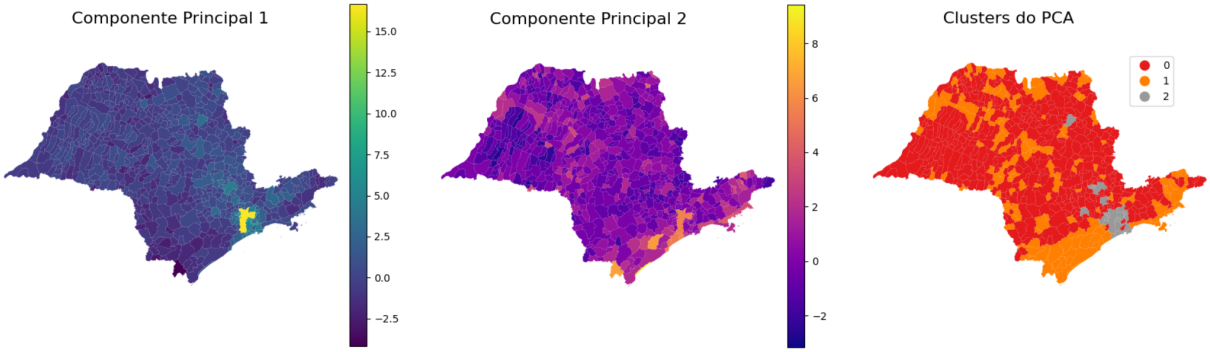
Figura 22 – PCA - municípios com base nos indicadores criminais e sociais



Fonte: Elaborada pela autora

O gráfico de dispersão, Figura 22, mostra a distribuição dos municípios segundo PC1 e PC2, e apresenta aglomerações densas em torno da origem, com alguns *outliers* evidenciando perfis criminais atípicos. A aplicação do algoritmo de agrupamento *K-means* identificou três *clusters* distintos: *Cluster* 0 (verde): Municípios com baixos valores em ambas as componentes — sugerem cidades pequenas, com menor densidade, IDHM e taxas de criminalidade. *Cluster* 1 (laranja): Municípios intermediários, com valores moderados nos eixos — perfil misto. *Cluster* 2 (azul): Municípios com altos valores principalmente em PC1 — áreas urbanas e densas, como a Região Metropolitana de São Paulo.

Figura 23 – Mapas coropléticos das componentes principais e clusters



Fonte: Elaborada pela autora

Os mapas coropléticos das componentes principais reforçam a heterogeneidade espacial. No mapa PC1 (componente urbano-criminal), os maiores valores se concentram na região metropolitana de São Paulo, evidenciando a intensidade da criminalidade patrimonial e da urbanização. Cidades do interior apresentam valores baixos. No mapa PC2 (componente de violência interpessoal), a distribuição é mais espalhada, com municípios no litoral, sudoeste e extremo oeste com valores mais elevados. Isso mostra que a violência letal e interpessoal não está

restrita à metrópole. E o mapa dos *Cluster*, mostra que o *Cluster* 0 (vermelho) domina grande parte do interior — municípios menos densos e com menos criminalidade, o *Cluster* 1 (laranja) em cidades de médio porte com criminalidade considerável, e *Cluster* 2 (cinza) concentra-se na Grande São Paulo, litoral e regiões com maior complexidade urbana.

Para identificar padrões regionais entre os municípios do estado de São Paulo, foi utilizada a técnica de análise de agrupamento *K-means*, com objetivo de classificar os municípios em grupos homogêneos quanto a suas características criminais. As variáveis utilizadas nesta análise foram as taxas dos nove tipos de crimes por 100 mil habitantes. Neste caso, optou-se por trabalhar com quatro clusters (k = 4), com base em uma análise exploratória prévia que considerou tanto a interpretação dos grupos formados quanto a coerência geográfica entre os municípios. Após o treinamento do modelo, cada município foi alocado em um dos grupos, Figura 24, conforme a similaridade de seus indicadores criminais, de acordo com a Tabela 5.

Conforme discutido anteriormente, o método de agrupamento *K-means* foi inicialmente aplicado em conjunto com a Análise de Componentes Principais (PCA), com o objetivo de explorar padrões de agrupamento em uma representação reduzida dos dados, destacando as principais fontes de variação capturadas pelo PCA. Na etapa atual, o *K-means* é aplicado diretamente sobre o conjunto completo de variáveis criminais, sem redução dimensional, visando identificar perfis distintos de municípios com base nas características originais das variáveis analisadas.

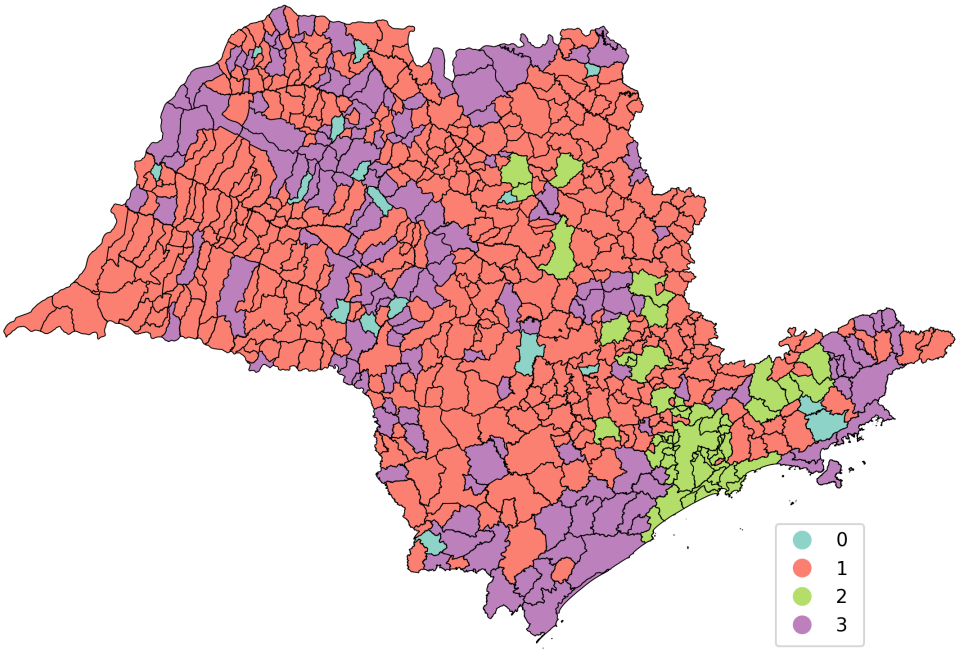
Tabela 5 – Centros dos clusters (médias padronizadas das variáveis)

Taxas	Cluster 0	Cluster 1	Cluster 2	Cluster 3
homicidio_doloso	0.570888	-0.317280	0.286432	0.673601
tentativa_de_homicidio	0.659196	-0.349972	-0.248098	0.927868
lesao_corporal_seguida_de_morte	4.898268	-0.154558	0.011281	-0.123096
latrocinio	0.208137	-0.019460	0.181277	-0.031765
estupro	0.475084	-0.241860	-0.630760	0.792250
homicidio_culposo	0.386410	-0.102781	-0.711514	0.464769
lesao_corporal	-0.197067	-0.190322	-0.684278	0.747111
roubo	-0.382346	-0.285862	2.667545	-0.100386
furto	-0.035990	-0.317583	1.007535	0.498911

Fonte: Elaborada pela autora

Com as médias padronizadas das variáveis por cada *cluster*, podemos observar que no *cluster* 0, concentram-se cidades com elevadas taxas de homicídio doloso, tentativa de homicídio e, principalmente, um valor muito alto para lesão corporal seguida de morte. O padrão sugere locais com conflitos violentos e letalidade acima da média, embora sem destaque para crimes patrimoniais. Já o *cluster* 1 agrupa municípios cujas taxas criminais estão, em geral, abaixo da média estadual. Esses municípios podem representar regiões mais seguras ou cidades de pequeno porte, onde os registros criminais são menos expressivos. O *cluster* 2 apresenta municípios com altas taxas de roubo e furto, características comuns em áreas urbanas, regiões comerciais ou polos turísticos. Em contrapartida, os índices de violência letal e interpessoal permanecem

Figura 24 – *Clusters* dos municípios por perfil criminal



Fonte: Elaborada pela autora

baixos. O *cluster* 3 destaca-se pelos municípios com altas taxas de homicídio doloso, tentativa de homicídio, lesão corporal e estupro, refletindo contextos de violência interpessoal e crimes contra a pessoa.

4.6 ESTUDO DE CASO

Durante a realização desse trabalho, foram realizadas dezenas de análises, uma delas chamou a atenção. A Tabela 6 mostra os municípios com maiores valores médios de taxa de furto, onde a maioria pertence a regiões litorâneas ou metropolitanas, e que já apresentavam altos índices em outras tabelas, entretanto, Barretos destacou-se como um caso atípico, aparecendo entre os dez municípios com maiores taxas de furto, apesar de não compartilhar essas mesmas características territoriais.

Para compreender melhor esse comportamento, foi realizada uma análise específica das ocorrências de furtos em Barretos ao longo do período de 2017 a 2024. A Figura 25 ilustra a evolução mensal das ocorrências por ano. Nota-se um padrão relativamente estável na maior parte do período, mas com um aumento expressivo e recorrente no mês de agosto.

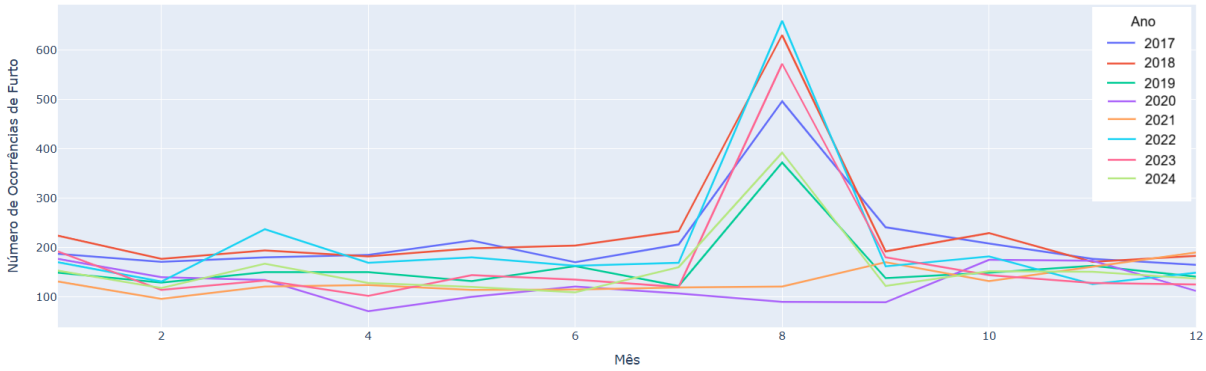
O comportamento observado coincide com a realização da Festa do Peão de Boiadeiro de Barretos, conhecida nacionalmente como Barretão, que atrai milhares de visitantes de diversas partes do país e até do exterior. O grande fluxo de turistas nesse período eleva significativamente a circulação de pessoas na cidade, aumentando a exposição a crimes patrimoniais, como os furtos. Os dois únicos anos que apresentam comportamentos diferentes em agosto, se mantendo constante em todos os meses, foram 2020 e 2021, anos que, devido à pandemia, não houve festa.

Tabela 6 – Municípios com maiores valores médios da taxa de furto

Município	Taxa de Furto
Ilha Comprida	5552.312
Mongaguá	3238.914
Itanhaém	2696.128
Peruíbe	2649.495
São Paulo	2074.769
Bertioga	1963.177
Praia Grande	1881.693
Barretos	1745.087
Lins	1731.154
Jeriquara	1727.792

Fonte: Elaborada pela autora

Figura 25 – Evolução mensal de ocorrências de furtos em Barretos por ano



Fonte: Elaborada pela autora

Assim, a sazonalidade observada não se explica apenas por fatores demográficos ou socioeconômicos fixos, mas também por eventos sazonais de grande porte que alteram temporariamente a dinâmica social e criminal do município.

Esse estudo de caso evidencia a importância de considerar eventos locais específicos na análise dos padrões criminais. Enquanto os índices médios apontam tendências gerais, análises direcionadas revelam peculiaridades que contribuem para uma compreensão mais completa da criminalidade. No caso de Barretos, a relação direta entre o aumento dos furtos e a realização de um evento festivo reforça a necessidade de políticas públicas de segurança voltadas à gestão de grandes eventos, com estratégias preventivas que considerem o impacto temporário, porém significativo, na dinâmica criminal do município.

5 CONCLUSÃO

Esta dissertação teve como objetivo principal analisar a evolução temporal e espacial dos crimes nos municípios do estado de São Paulo, buscando identificar padrões relevantes para a compreensão da dinâmica criminal. A análise exploratória inicial permitiu caracterizar os dados de forma abrangente, destacando a elevada heterogeneidade entre os municípios.

A avaliação temporal revelou tendências e sazonalidades distintas para os diferentes crimes, além de apontar alterações significativas durante o período da pandemia de COVID-19, salientando como eventos sociais impactam diretamente os índices criminais. Crimes patrimoniais possuem sazonalidade, acontecem mais em meses específicos, certamente em datas festivas ou com maior circulação de pessoas, enquanto que crimes contra a vida não apresentam esse tipo de padrão tão claramente. Comparando os anos de 2017 e 2024, crimes como estupro, homicídio culposo e lesão corporal seguida de morte, tiveram aumento significativo, enquanto que roubo, homicídio doloso e latrocínio tiveram queda. Mostrando mais uma vez que não tem um padrão nem entre categorias de crimes, o que dificulta ações preventivas.

As análises espaciais, confirmaram a existência de dependência espacial nas ocorrências criminais, mostrando que a criminalidade em um município tende a influenciar os municípios vizinhos e evidenciando a concentração espacial e a persistência de certos tipos de crimes em regiões específicas, como roubo e furto em regiões litorâneas e Grande São Paulo, onde o fluxo de turistas, por exemplo, é maior, e essa população flutuante interfere muito nas taxas por não ser quantificada. Mas outros tipos de crime, com menor ocorrência, como lesão corporal seguida de morte, não tem uma correlação com o local, não tem influência de vizinhos e se apresenta de forma espalhada no estado. Os métodos multivariados aplicados ampliaram a compreensão das inter-relações entre os indicadores e permitiram a identificação de agrupamentos de municípios com perfis criminais semelhantes. Mostrando que municípios no litoral e no interior do estado com mesmo porte populacional possuem perfis diferentes de criminalidade, reforçando a importância de tratar os perfis criminais de forma segmentada e específica.

Por fim, a investigação das correlações trouxe que crimes de maiores incidências como furto e roubo, tem uma correlação alta com população, já PIB e IDHM mostrou pouca correlação com os crimes, sugerindo que fatores estruturais como desenvolvimento humano e desigualdade podem influenciar, mas não explicam os padrões observados e indicam que a criminalidade está associada a uma combinação de fatores sociais, demográficos e territoriais de uma forma muito mais complexa.

Os resultados alcançados contribuem para o entendimento mais aprofundado da criminalidade em São Paulo, oferecendo auxílio para políticas públicas mais direcionadas e estratégias de segurança mais eficazes com abordagens específicas para prevenção e combate à criminalidade, baseadas em evidências empíricas.

Como perspectivas futuras, este trabalho pode ser ampliado por meio do uso de modelos

preditivos, utilizando métodos estatísticos e de aprendizado de máquina para antecipar ocorrências criminais em nível municipal, contribuindo para sistemas de alerta antecipado e apoio à tomada de decisão. Outra vertente promissora é a incorporação de novas bases de dados, como indicadores de mobilidade urbana e variáveis socioeconômicas atualizadas, que permitiriam análises mais abrangentes e contextualizadas. Do ponto de vista metodológico, o uso de modelos espaciais avançados, como regressões ponderadas geograficamente, possibilitaria investigar de forma mais refinada a heterogeneidade territorial da criminalidade. Por fim, investigações voltadas ao impacto de políticas públicas e de eventos sazonais de grande porte, como a Festa de Peão em Barretos, ampliam o potencial de estudos futuros, reforçando a relevância da integração entre ciência de dados e segurança pública.

REFERÊNCIAS

ALMEIDA, M. A. S. de. **Análise exploratória e modelo explicativo da criminalidade no estado de São Paulo: interação espacial (2001)**. 2007. Dissertação (Mestrado em Economia) — Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 2007.

ANDRADE, S. R. C. de. **Análise exploratória sobre a criminalidade utilizando dados espaciais.** 2022. Trabalho de Conclusão de Curso (Especialização em Ciência de Dados) — Universidade Tecnológica Federal do Paraná, Dois Vizinhos, 2022. Disponível em: https://repositorio.utfpr.edu.br/jspui/handle/1/32385. Acesso em: 4 maio 2025.

ANDRIENKO, G.; ANDRIENKO, N.; SAVINOV, A. Choropleth maps: classifi cation revisited. **Proc. 20th International Cartographic Conference - ICA'2001**, Beijing, China, p. 1209–1219, 2001.

ANSELIN, L. **Spatial econometrics: methods and models**. Dordrecht, Holanda: Springer, 1988.

ANSELIN, L. Local indicators of spatial association—lisa. **Geographical Analysis**, Columbus, OH, v. 27, n. 2, p. 93–115, 1995.

ANSELIN, L. Spatial econometrics. In: MANFRED M. FISCHER, PETER NIJKAMP. **Handbook of regional science**: linguagem & comunicação. 2. ed. Berlin, Heidelberg: Springer, 2019. p. 1–19.

ARAúJO, M. V. A. **Um modelo espaço-temporal para a criminalidade nos bairros de Fortaleza: a influência de fatores comuns e o efeito vizinhança**. 2018. Dissertação (Mestrado em Economia) — Faculdade de Economia, Administração, Atuária e Contabilidade, Universidade Federal do Ceará, Fortaleza, 2018.

BASTOS, J. L. D.; DUQUIA, R. P. Medidas de dispersão: os valores estão próximos entre si ou variam muito? **Scientia Medica** — comunicação, saúde e educação, Porto Alegre, v. 17, n. 1, p. 40–44, 2007.

BERG, R. A. V. D. et al. A practical guideline for large-scale data analysis using log transformation. **PLoS ONE**, v. 16, n. 1, p. e0246464, 2021. Disponível em: https://doi.org/10.1371/journal.pone.0246464.

BUTT, U. M. et al. Spatio-temporal crime hotspot detection and prediction: a systematic literature review. **IEEE Access**, v. 8, p. 166553–166574, 2020. DOI: http://dx.doi.org/10.1109/access.2020.3022808.

CLEVELAND, W. S. **Visualizing data**. Summit, New Jersey: Hobart Press, 1993.

CONOVER, W. J. **Practical Nonparametric Statistics**. 3rd. ed. [S.l.]: John Wiley & Sons, 1999.

COURTNEY, M. B. Exploratory data analysis in schools: A logic model to guide implementation. **International Journal of Education Policy and Leadership**, v. 17, n. 4, p. 14f, 2021. DOI: 10.22230/ijepl.2021v17n4a1041.

DAVOGLIO, G. R. **Roubo e furto no estado de São Paulo em 2016: análise espacial e variáveis explicativas**. 2019. Dissertação (Mestrado em Economia Regional) — Universidade Estadual de Londrina, Londrina, 2019.

EVERITT, B. S. et al. **Cluster analysis**. Chichester: John Wiley & Sons, 2011.

FALCO, J. G. **Estatística aplicada**. Cuibá: EdUFMT, 2008.

FEIJOO, A. M. L. C. de. **A pesquisa e a estatística na psicologia e na educação**. Rio de Janeiro: Centro Edelstein de Pesquisas Sociais, 2010.

FILHO, J. G. T. **Teoria do crime: evolução histórica**. 2023. Disponível em: https://www.jusbrasil.com.br/artigos/teoria-do-crime-evolucao-historica/1563376528. Acesso em: 25 abr. 2025.

FRANCIA, T. A. **A criminalidade no estado de São Paulo: uma análise espacial da relação do mercado de drogas ilícitas com os crimes de roubo e homicídio doloso.** 2024. Trabalho de Conclusão de Curso (Graduação em Ciências Econômicas) — Universidade Federal de São Paulo, Osasco, 2024. Disponível em: https://repositorio.unifesp.br/items/5119ab0f-7d90-4154-a6ed-6fd2ef88668d. Acesso em: 28 abr. 2025.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. **Anuário Brasileiro de Segurança Pública**. 3. ed. Rio de Janeiro, 2023. Disponível em: https://forumseguranca.org.br/wp-content/uploads/2023/07/anuario-2023.pdf. Acesso em: 15 abr. 2025.

GARCIA, G. et al. Crimanalyzer: Understanding crime patterns in são paulo. **IEEE Transactions on Visualization and Computer Graphics**, v. 27, n. 4, p. 2313–2328, 2021.

GAULEZ, M. P.; MACIEL, V. F. **Determinantes da Criminalidade no Estado de São Paulo: Uma análise Espacial de dados em Cross-Section**. [S.l.], 2016. Disponível em: https://EconPapers.repec.org/RePEc:anp:en2015:201. Acesso em: 20 abr. 2025.

GETIS, A.; ORD, J. K. The analysis of spatial association by use of distance statistics. **Geographical Analysis**, Columbus, OH, v. 24, n. 3, p. 189–206, 1992.

GOODCHILD, M. F. **Spatial Autocorrelation**. Norwich, UK: Geo Books, 1986.

Google Research. **Google Colaboratory**. 2024. Disponível em: https://colab.research.google.com. Acesso em: 24 jul. 2025.

GUIMARãES, F. T.; BECKER, K. L. A criminalidade no rio grande do sul: Análise exploratória de dados espaciais para os anos de 2002, 2010 e 2018. **Revista Economia Ensaios**, Uberlândia, v. 36, n. 2, p. 189–206, 2021. DOI: https://doi.org/10.14393/REE-v36n2a2021-53792. Disponível em: https://seer.ufu.br/index.php/revistaeconomiaensaios/article/view/53792. Acesso em: 8 abr. 2025.

HAIR, J. F. et al. **Multivariate Data Analysis**. Upper Saddle River, NJ: Pearson Prentice Hall, 2009.

HINTZE, J. L.; NELSON, R. D. Violin plots: A box plot-density trace synergism. **The American Statistician**, ASA Website, v. 52, n. 2, p. 181–184, 1998. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480559. Acesso em: 29 abr. 2025.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Previsão: princípios e prática**. Melbourne, Austrália: OTexts, 2021. Disponível em: https://otexts.com/fpp2/. Acesso em: 01 mai. 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Demográfico**. Rio de Janeiro, 2023. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/saude/22827-censo-demografico-2022.html?=&t=downloads. Acesso em: 11 fev. 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Malha Municipal**. Rio de Janeiro, 2024. Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html. Acesso em: 20 mar 2025.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: A review and recent developments. **Philosophical Transactions of the Royal Society A**, v. 374, n. 2065, p. 20150202, 2016.

MCKINNEY, W. **Python para Análise de Dados: Tratamento de Dados com Pandas, NumPy e IPython**. 1. ed. São Paulo: Novatec Editora, 2018. Cópia do arquivo PDF não-oficial obtida em GitHub (Acesso em [Dia Mês Ano do Acesso]). ISBN 978-8575226476. Disponível em: https://github.com/Ediandrofc/livros-datascience/blob/master/Python%20para%20analise%20de%20dados%20-%20Wes%20McKinney.pdf.

MORAN, P. A. P. Notes on continuous stochastic phenomena. **Biometrika**, London, v. 37, n. 1/2, p. 17–23, 1950.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. 4. ed. São Paulo: Blucher, 2017.

MUKAKA, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. **Malawi Medical Journal**, v. 24, n. 3, p. 69–71, 2012.

MüLLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. 1. ed. Sebastopol, CA: O'Reilly Media, Inc., 2016. ISBN 978-1449369415.

NERY, M.; ADORNO, S. Crime e violências em são paulo: retrospectiva teórico-metodológica, avanços, limites e perspectivas futuras. **Cadernos Metrópole**, São Paulo, v. 24, n. 54, p. 409–433, 2022. Disponível em: https://www.scielo.br/j/cm/a/W4wbLBTYnNdKLVR4CVH3FSS/. Acesso em: 3 mai. 2025.

NERY, M. et al. **TensorAnalyzer: Identification of Urban Patterns in Big Cities Using Non-Negative Tensor Factorization**. 2022. ArXiv preprint arXiv:2210.02623. Disponível em: https://arxiv.org/pdf/2210.02623. Acesso em: Acesso em: 3 maio 2025.

PAZ, I. S. da. **Índice de Exposição aos Crimes Violentos: 2021 e 2022**. 5. ed. São Paulo, 2023. Disponível em: https://soudapaz.org/documentos/iecv-indice-de-exposicao-aos-crimes-violentos-2021-2022/.

PRADO, K. H. de J. **Data Science Aplicada à Análise Criminal Baseada nos Dados Abertos Governamentais do Brasil**. 2019. Dissertação (Mestrado em Informática) — Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2019.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking**. Sebastopol, CA: O'Reilly Media, 2013.

REY, S. J.; ANSELIN, L. Pysal: A python library of spatial analytical methods. **The Review of Regional Studies**, v. 37, n. 1, p. 5–27, 2007.

SILVA, F. C. C. da. Visualização de dados: passado, presente e futuro. **LIINC em Revista**, Rio de Janeiro, RJ, v. 15, n. 2, p. 205–223, 2019. Disponível em: https://revista.ibict.br/liinc/article/view/4586. Acesso em: 27 mar. 2025.

SILVA, F. R. da; RAMOS, E. M. L. S. **Categorização dos Crimes Violentos no Brasil**. Belém, 2024. Disponível em: http://educapes.capes.gov.br/handle/capes/741218. Acesso em: 10 abr. 2025.

SSP-SP, S. d. S. P. d. S. P. **Ocorrências Policiais registradas por Mês - Dados abertos**. 2024. Disponível em: https://www.ssp.sp.gov.br/estatistica/dados-mensais. Acesso em: 10 fev. 2025.

TABACHNICK, B. G.; FIDELL, L. S. **Using Multivariate Statistics**. 7. ed. Harlow, UK: Pearson Education, 2019.

TUKEY, J. W. **Exploratory Data Analysis**. Reading, MA: Addison-Wesley, 1977.

WILKINSON, L. **The Grammar of Graphics**. 2. ed. New York: Springer Science & Business Media, 2005.

ZANABRIA, G. G. **Visual Crime Pattern Analysis.** Orientador: Luis Gustavo Nonato. 2021. 134 p. Tese (Doutorado em Ciências de Computação e Matemática Computacional) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-09042021-161411/pt-br.php. Acesso em: 3 maio 2025.

ZANABRIA, G. G. et al. Cripav: Street-level crime patterns analysis and visualization. **IEEE Transactions on Visualization and Computer Graphics**, v. 28, n. 12, p. 4000–4015, 2022. Disponível em: https://ieeexplore.ieee.org/document/9536407. Acesso em: 4 maio 2025.