Detetives de Dados - Resultados da Pesquisa

Anderson Nogueira Silva - 2126516
Breno Moura de Abreu - 1561286
Heitor Derder Trevisol - 1611810

[1]Departamento Acadêmico de Informática (DAINF) –
Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de Setembro, 3165 – 80230-901 – Curitiba – PR – Brasil

1. Introdução

O tema deste trabalho é o estudo de tendências na criminalidade na cidade de Curitiba com análises regionais e temporais. O objetivo principal da equipe é entender como está situada a distribuição de crimes e ocorrências policiais, levando em consideração a região, o tipo do crime e dados temporais registrados. Além disso, também temos o objetivo de verificarmos as tendências de criminalidade em Curitiba, permitindo estimar o estado futuro das ocorrências.

2. Processamento dos Dados

Para responder às perguntas de pesquisa foi utilizada a base de dados da SiGes Guarda [e Trânsito 2022] que contém dados sobre ocorrências em Curitiba entre os anos de 2009 e 2022; contém a data e hora, bairro e tipo da ocorrência. Também foi utilizado dados do IBGE [IBGE 2022] para adquirir os dados da população estimada de Curitiba em diferentes anos. A aplicação dos testes de significância estatística, optou-se por utilizar os conjuntos de dados de maneira que o número total de ocorrências ao longo de 2009 até 2022, fosse distribuído, para cada hipótese, da seguinte maneira

- para cada dia da semana (segunda, terça, quarta, etc);
- para cada mês do ano (janeiro, fevereiro, abril, maio, etc);
- para cada hora do dia (0h, 1h, 2h, . . . , 10h, 11h, 12h, . . . , 21h, 22h, 23h);

Para a aplicação dos modelos de tendência criminal pelos anos os dados foram agrupados em um período mensal utilizando ocorrências de natureza criminosa, os dados foram então testados para encontrar uma possível sazonalidade, a qual foi encontrada no período de 12 meses, os dados foram então reestruturados com o uso de média móvel com o peso 7xf(12) com objetivo de criar um conjunto de dados dessazonalizado para então aplicar Análise de Série Temporal (AST) nos dados com e sem a sazonalidade.

A preparação de dados para permitir a aplicação dos modelos de detecção de outliers se deu pela inclusão apenas de ocorrências de natureza criminal, da inclusão de uma coluna que indica o período do dia da ocorrência (manhã, tarde, noite e madrugada), do mapeamento dos bairros para suas regiões (9 regiões no total [5]), da inclusão da coluna que indica a semana do mês (dividido em 4 semanas por mês), do mapeamento do dia da semana para dia útil ou fim de semana, da normalização do número de ocorrências por grupo pela população estimada do ano em questão. Também foi utilizado o One-Hot-Encoding para codificar as variáveis categóricas criando novas colunas binárias para cada nova categoria.

3. Resultados

Nesta seção estão descritas as três perguntas direcionadoras das pesquisas, as hipóteses estabelecidas, bem como a conclusão obtida por parte das análises realizadas.

3.1. Existe alguma relação entre a data, hora, mês, dia da semana e as ocorrências?

As hipóteses e as análises realizadas sobre esta pergunta estão descritas a seguir

A. Nos fins de semana há mais ocorrências comparado com os demais dias da semana;
   - Teste U de MANN-WHITNEY apontou um p-value equivalente a 0.047619, como p-value $< 0,05$, portanto há diferenças estatisticamente significativas, apontando um maior número de ocorrências aos finais de semana.

B. Nos meses de inverno (Junho-Setembro), há menos ocorrências em geral que nos demais meses do ano;
   - Teste T independente apontou um p-value equivalente a 0.798806, como p-value $> 0,05$, portanto não há diferenças estatisticamente significativas entre as médias de ocorrências em meses de inverno com relação a outros meses do ano.

C. Há mais ocorrências no período noturno;
   - Teste T independente devolveu um p-value equivalente a 0.952146 e, como p-value $> 0,05$, além disso, p-value ¿ 0.95 foi apontado diferença estatisticamente significativa entre a média de ocorrências no período diurno ser maior do que no período noturno, o oposto da hipótese estabelecida.

D. Há mais ocorrências criminosas no período noturno.
   - Teste T independente retornou um p-value equivalente a 0.626015 e, como p-value $> 0,05$, não há diferença estatística significativa entre o número de ocorrências criminosas no período noturno, com relação ao período diurno.

3.2. Qual é a tendência para o futuro em relação à criminalidade em Curitiba?

As hipóteses estabelecidas foram

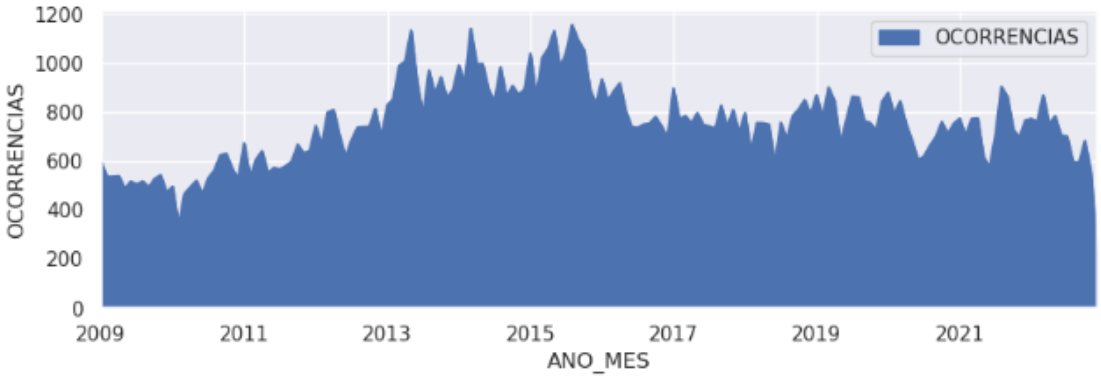
A. Com o aumento da população há aumento de ocorrências;
B. A cidade está ficando mais violenta com o passar dos anos.

3.2.1. Normalização

Os dados utilizados para análises de tendência foram primeiramente normalizados para analisar o efeito do aumento da população estimada sobre as ocorrências criminosas no mesmo período. A figura 1 apresenta a distribuição ao longo dos anos de 2009 a 2022.

Como pode ser facilmente observado o número de ocorrências criminosas não aparenta ser afetado pelo crescimento populacional da cidade. Esses resultados são corroborados pelas análises temporais feitas a partir dos dados.

Figura 1. Visualização das ocorrências normalizadas pela população pelo tempo.



Fonte: Autores.

3.2.2. ANÁLISE DE SÉRIE TEMPORAL (AST)

Dois modelos foram utilizados para realizar estas análises, um deles utilizando os dados brutos com o efeito da sazonalidade anual encontrada, e o outro utilizando os dados que passaram pelo processo de dessazonalização via aplicação de média móvel. A escolha de aplicar esses dois modelos se deu pela análise do p-value encontrado pelo método Augmented Dickey Fuller (ADF), o qual encontrou um valor de 0,2631 para os dados brutos e 8,545*10-5 para os valores trabalhados.

Figura 2. Tendência temporal das ocorrências criminosas.

Fonte: Autores.

Como visto na figura 2 a hipótese sobre a tendência criminal de que a cidade está ficando mais violenta não é confirmada pelos resultados encontrados nestas análises.

3.3. Onde é possível observar anomalias relacionadas à criminalidade em Curitiba?

As hipóteses estabelecidas para essa pergunta foram

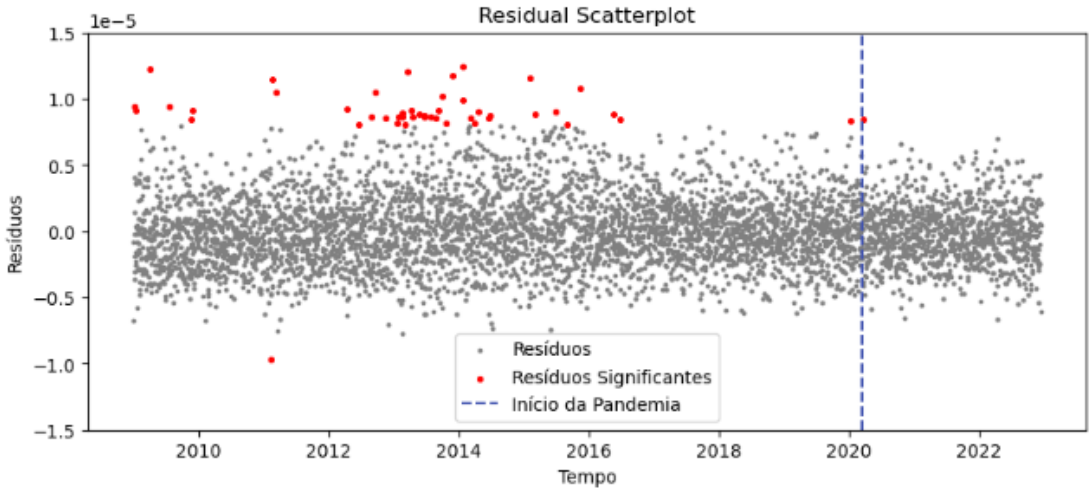
A. No período logo após o início da pandemia de COVID-19 houveram valores anômalos relacionados com a criminalidade de Curitiba;
B. Instâncias anômalas são mais prováveis de aparecerem em regiões específicas.

3.3.1. ARIMA

A análise realizada a partir do ARIMA foi a visualização dos resíduos gerados pela aplicação do modelo utilizando a quantidade de ocorrências criminais por dia. O objetivo é encontrar dias que apresentaram um número anômalo de ocorrências criminais.

A Figura 3 apresenta os resultados encontrados, demonstrando que não houve anomalias consideráveis a partir de Março de 2020.

Figura 3. Visualização dos valores residuais por dia. Pontos vermelhos indicam anomalias.



Fonte: Autores.

Como é possível observar, os pontos mais distantes do seu valor previsto são mais visíveis entre os anos de 2009 e 2017. Após o início da pandemia, não há valores consideravelmente distantes dos previstos que indiquem um número anômalo de ocorrências naquele ano em específico.

3.3.2. SOM

Para aplicação do modelo SOM, os dados foram mapeados de acordo com

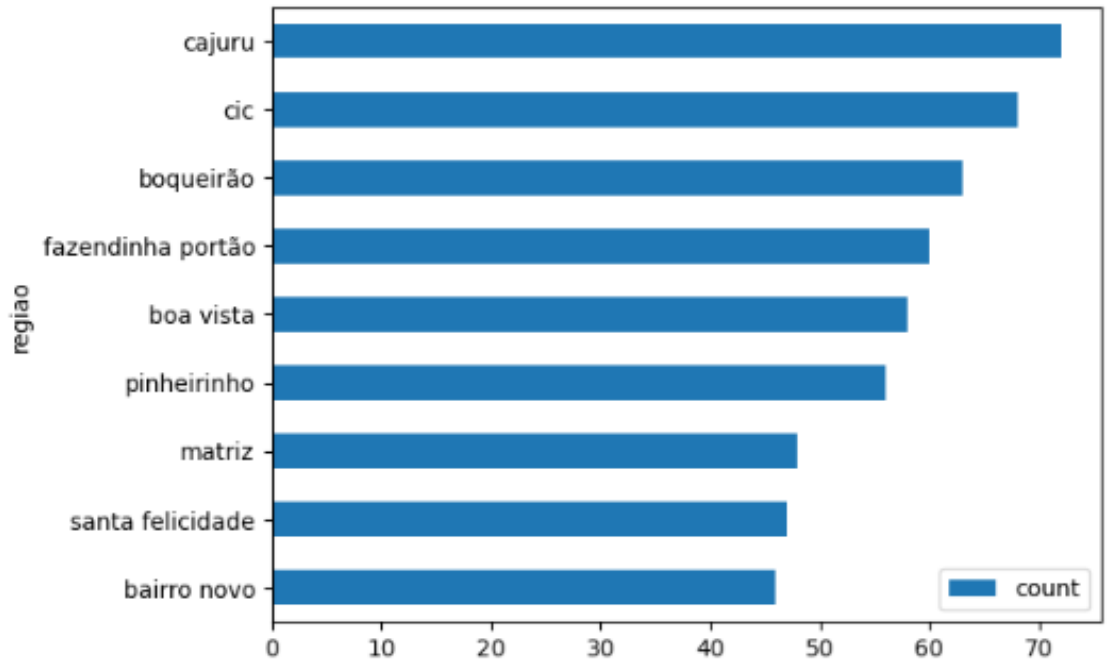
- região (9 regiões);
- período do dia (madrugada; manhã; tarde e noite);
- semana do mês (até 4 semanas);
- tipo do dia da semana (dia útil e finais de semana i.e. sábado e domingo).

Ainda, os dados foram agrupados pelos valores temporais, região e ocorrência. Uma nova coluna foi criada contendo o número da ocorrência para cada agrupamento pelo número estimado de habitantes no ano em questão.

Com os dados preparados para conter apenas valores numéricos e normalizados, o modelo foi aplicado e as distâncias foram calculadas. Com os dados ordenados, foi selecionado 1% das instâncias mais distantes para serem analisadas como sendo outliers.

A figura 4 apresenta os resultados em relação à distribuição das anomalias por região de Curitiba.

Figura 4. Distribuição de anomalias por região.



Fonte: Autores.

Como é possível observar, não há um valor acentuado para uma região específica o que indica que a segunda hipótese - que afirma que instâncias anômalas são mais prováveis de aparecer em regiões específicas - é falsa. O que demonstra que, apesar de algumas regiões apresentarem um comportamento mais previsível que outras, em geral as anomalias não estão relacionadas com este fator.

4. Limitações e Trabalhos Futuros

Um desafio para a aplicação dos modelos está na compreensão dos parâmetros, entendendo como sua alteração modifica os resultados. É necessário realizar diversos testes para encontrar a melhor combinação de parâmetros e ainda assim é difícil dizer se os resultados encontrados são eficazes. Para alguns casos, por se tratar de aprendizado não-supervisionado, não é possível ter certeza se os valores encontrados estão de acordo com a realidade ou se foram o resultado de um modelo mal-construído. É necessário analisar os resultados a cada execução nova e é difícil ter uma boa ideia de como alterar os argumentos para que o modelo seja mais eficiente.

Como trabalho futuro pode se citar o estudo de clusterização dos bairros da cidade relacionados com os tipos de crime, estudo este previsto no nosso projeto, mas que não foi concluído pela equipe até a data final de entrega.

Referências

do Trabalho de Curitiba, O. (2016). Estudo temÁtico 2: Perfil demográfico e socioeconômico dos bairros agregados de curitiba. acesso em 10 maio 2023.

e Trânsito, D. S. (2022). Base de dados sigesguarda.

GeoPandas Development Team (2021). GeoPandas: Python tools for geographic data.

IBGE (2022). Panorama cidade de curitiba, paraná. acesso em 12 maio 2023.

IPPUC (2018). Instituto de pesquisa e planejamento urbano de curitiba - dados geográficos. acesso em 15 abril 2023.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. In Loizides, F. and Schmidt, B., editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press.

Matplotlib Development Team (2021). Matplotlib: A 2D graphics environment.

Pandas Development Team (2021). Pandas: Powerful data analysis tools for Python.

Prefeitura Municipal de Curitiba (acesso em 16 abril 2023). Portal de dados abertos de curitiba.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Wikipedia. Lisa de bairros de curitiba. acesso em 12 maio 2023.