UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Gustavo Freitas Cunha 2020054498
Gabriel Martins Juarez 2018046530

Trabalho Prático
Introdução à Ciência dos Dados

Belo Horizonte
Julho de 2023

1. INTRODUÇÃO

Segundo o Fórum Brasileiro de Segurança Pública, em 2020, ao comparar as taxas de Mortes Violentas Intencionais por zonas de ocupação, verifica-se que o conjunto de municípios com as maiores taxas são os sob pressão de desmatamento (37,1 por 100 mil habitantes), em segundo lugar, os municípios desmatados (34,6), os municípios não florestais com taxa de 29,7, e, por fim, os municípios florestais apresentam a menor taxa de letalidade violência, com 24,9 por 100 mil.

Neste trabalho, queremos verificar se, de fato, há correlação entre a taxa de desmatamento e a taxa de homicídios de cidades da Amazônia legal ao longo dos anos, e como se dá essa correlação. Para isso, responderemos a quatro perguntas:

- É possível prever a taxa de homicídios de uma cidade da Amazônia Legal com base em sua taxa de desmatamento?

- O quão boa é a correlação entre a taxa de homicídios e a taxa de desmatamento de cidades da Amazônia Legal?

- É possível classificar uma cidade da Amazônia Legal como perigosa ou não segundo sua taxa de desmatamento?

- Quais são os pares de cidades da Amazônia Legal com taxa de desmatamento semelhante e taxa de homicídios divergentes?

2. METODOLOGIA

Neste trabalho, montamos dois dataframes principais, através da junção e limpeza de vários dataframes que encontramos em sites governamentais de dados abertos. Fazendo isso, chegamos aos dataframes **df_desmatamento_amazonia_legal** e **df_homicidios_amazonia_legal**. Ambas as bases possuem um identificador do município (apenas município da Amazônia Legal), o ano de referência a taxa de desmatamento e taxa de homicídios, respectivamente. O frame de desmatamento possui dados de 2014 a 2021 e o dataframe de homicídios de 2014 a 2019.

Desse modo, nossa variável dependente é a taxa de homicídios e a variável independente é a taxa de desmatamento, isto é, **vamos tentar prever a taxa de homicídios com base na taxa de desmatamento de municípios da Amazônia Legal**.

3. ANÁLISE EXPLORATÓRIA

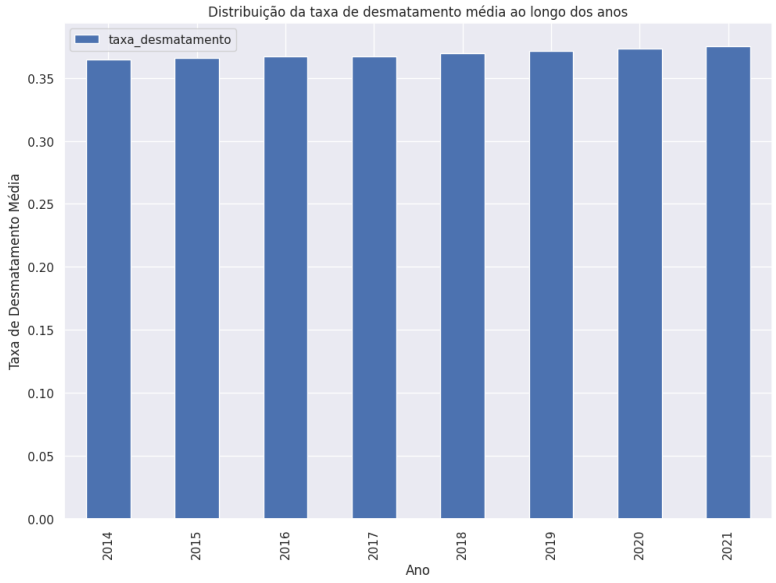
A base de dados utilizada neste estudo contém informações sobre taxas de homicídios, características geográficas e dados relacionados ao desmatamento em municípios da região da Amazônia Legal ao longo do período de 2014 a 2019. O conjunto de dados consiste em 4.560 registros, abrangendo uma ampla gama de atributos.

Ao analisar os dados, observamos que a taxa média de homicídios na região da Amazônia Legal foi de aproximadamente 0.12, com um desvio padrão de 0.10, indicando certa variabilidade nas taxas entre os municípios. Além disso, a base de dados oferece insights sobre a área total dos municípios, com uma média de 6668.48 e um desvio padrão de 13858.61. Essa informação é relevante para compreender a extensão geográfica dos municípios estudados.

No contexto do desmatamento, observa-se que a taxa média de desmatamento foi de 0.37, com

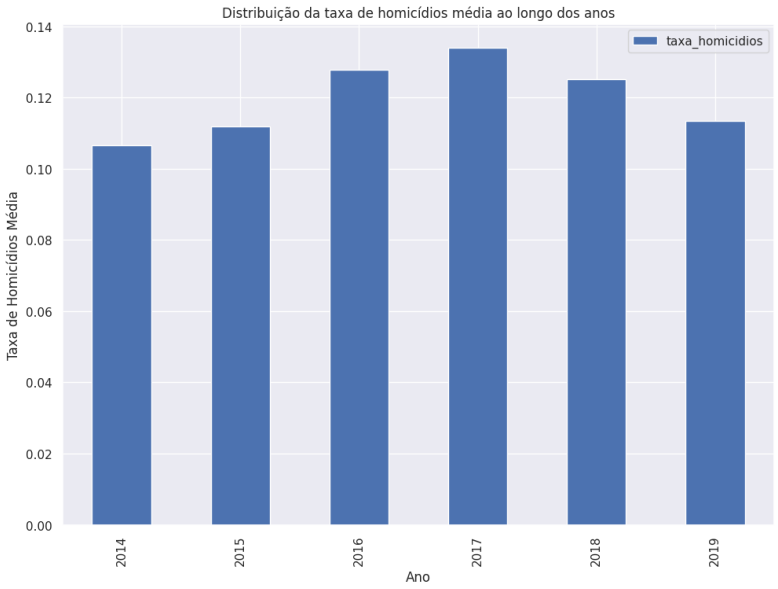
um desvio padrão de 0.33. Esses números sugerem uma variação considerável nas taxas de desmatamento entre os municípios.

- Gráfico 1:



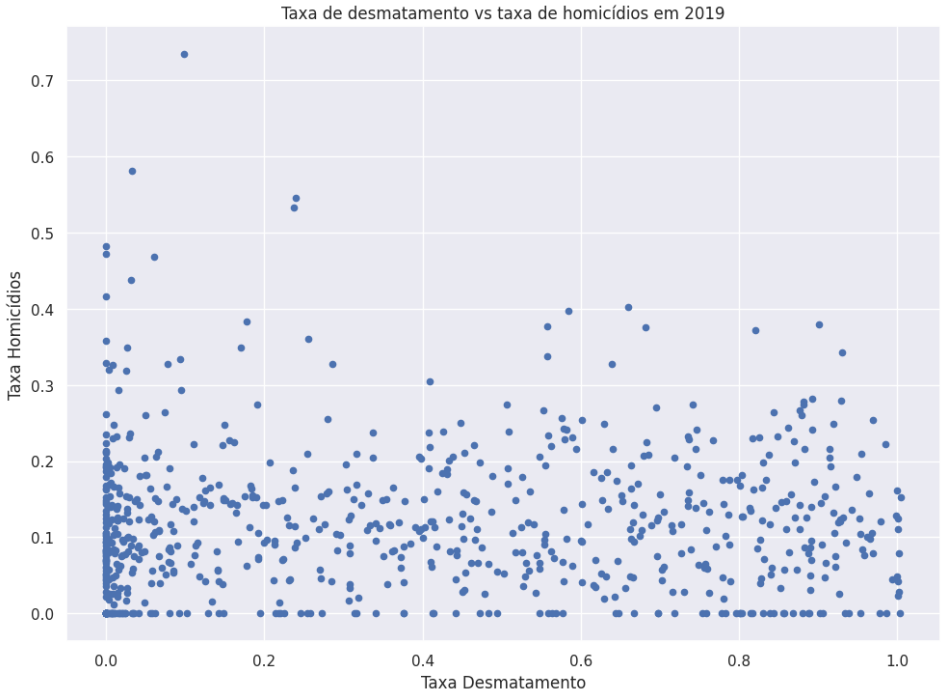
Neste histograma, plotamos a média da taxa de desmatamento de todos os municípios de 2014 a 2021. Nota-se que a taxa apresenta um leve crescimento, mas se mantém entre 0.35 e 0.40 no período analisado.

- Gráfico 2:



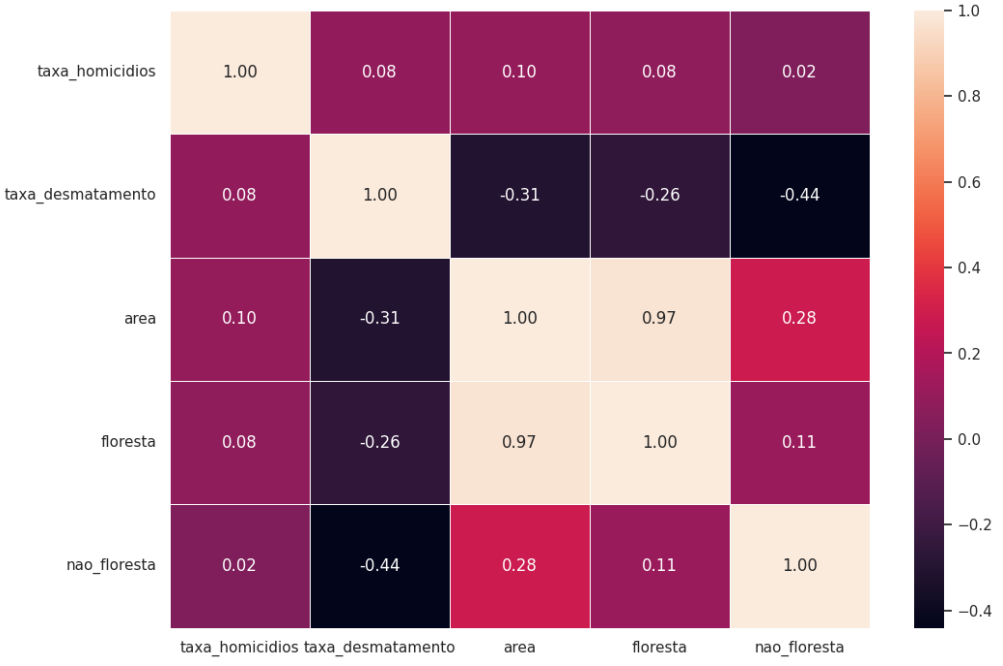
Neste histograma, plotamos a média da taxa de homicídios de todos os municípios de 2014 a 2019. Nota-se que a taxa apresenta um crescimento até o pico em 2017 e depois uma queda até 2019, mas se manteve entre 0.10 e 0.13.

- Gráfico 3:



Neste scatter plot, vemos como a variável independente e a variável resposta da nossa análise se comportam. De fato, parece que não há uma correlação muito grande entre elas.

- Gráfico 4:



Neste mapa de calor, plotamos a correlação entre os pares de variáveis mais relevantes para nossa análise: taxa de homicídios, taxa de desmatamento, área total do município, área florestal, e área não florestal.

4. MÉTODOS E MODELOS UTILIZADOS

4.1 Regressão Linear Simples

Para realizar a análise de regressão, coletamos dados que incluíam a taxa de homicídios e a taxa de desmatamento para cada município da região de estudo. Utilizamos a biblioteca Scikit-learn do Python, que oferece ferramentas para realizar análises estatísticas e ajustar modelos de regressão. O objetivo era determinar se havia uma relação linear entre as duas variáveis.

Utilizamos a regressão linear simples, que assume uma relação linear entre a variável dependente (taxa de homicídios) e a variável independente (taxa de desmatamento). Calculamos os coeficientes angular (slope) e linear (intercept) por meio da função polyfit do NumPy. Com base nesses coeficientes, geramos uma reta de regressão que representa a melhor estimativa da relação linear entre as variáveis.

Após ajustar o modelo de regressão, avaliamos sua qualidade utilizando o coeficiente de determinação R^2. O valor obtido foi de 0.007, o que indica que apenas 0.7% da variabilidade na taxa de homicídios é explicada pela taxa de desmatamento. Esse valor próximo de zero sugere que a regressão linear não é adequada para capturar a relação entre essas variáveis.

Além disso, analisamos as correlações de Pearson entre as variáveis taxa de homicídios e taxa de desmatamento. A correlação entre a taxa de homicídios e a taxa de desmatamento foi de 0.083674, um valor próximo de zero. Isso indica uma correlação muito fraca entre essas variáveis.

Os resultados da regressão construída e as previsões feitas serão discutidos nas próximas seções.

4.1 Intervalo de Confiança

O intervalo de confiança é calculado a partir dos dados amostrais e fornece uma faixa de valores plausíveis para os coeficientes da regressão populacional. Essa faixa leva em consideração a incerteza inerente à estimativa dos coeficientes com base em uma amostra limitada de dados.

No código, é utilizado o método conf_int do objeto results retornado pelo ajuste do modelo de regressão linear. Esse método calcula o intervalo de confiança para os coeficientes com base nas propriedades estatísticas do modelo ajustado.

O intervalo de confiança é definido pelo nível de confiança escolhido, que representa a probabilidade de que o intervalo contenha o verdadeiro valor do coeficiente populacional. No código, é utilizado um nível de confiança de 0.05 (ou 95% de confiança), o que significa que espera-se que os intervalos de confiança contenham os verdadeiros valores dos coeficientes com uma probabilidade de 95%.

4.2 Testes de Hipótese

- **Teste 1: É possível prever a taxa de homicídios de uma cidade da Amazônia Legal com base em sua taxa de desmatamento?**

Neste teste, é realizada uma regressão linear para avaliar a relação entre a taxa de desmatamento e a taxa de homicídios em uma cidade da Amazônia Legal. É ajustado um modelo de regressão linear utilizando a biblioteca statsmodels e é calculado o coeficiente de determinação (R^2) para medir o quão bem o modelo se ajusta aos dados. Em seguida, é realizado um teste de hipótese sobre o coeficiente de determinação, onde a hipótese nula (H0) afirma que não há relação entre as variáveis e a hipótese alternativa (H1) afirma que há uma relação significativa. O teste estatístico utilizado é o teste F, onde é comparado o modelo ajustado com um modelo nulo que não possui variáveis explicativas. Se o valor-p do teste F for menor que o nível de significância (geralmente 0,05),

rejeita-se a hipótese nula e conclui-se que há uma relação significativa entre a taxa de desmatamento e a taxa de homicídios.

- **Teste 2: O quão boa é a correlação entre a taxa de homicídios e a taxa de desmatamento de cidades da Amazônia Legal?**

Neste teste, é calculada a correlação de Pearson entre a taxa de desmatamento e a taxa de homicídios das cidades da Amazônia Legal. A correlação de Pearson é uma medida estatística que avalia a direção e a força da relação linear entre duas variáveis contínuas. O coeficiente de correlação varia de -1 a 1, onde valores próximos de -1 indicam uma correlação negativa perfeita, valores próximos de 1 indicam uma correlação positiva perfeita e valores próximos de 0 indicam uma correlação fraca ou nula. No teste de hipótese, a hipótese nula (H0) assume que não há correlação entre as variáveis, enquanto a hipótese alternativa (H1) assume que há uma correlação significativa. O valor-p é calculado e comparado com um nível de significância pré definido (geralmente 0,05). Se o valor-p for menor que o nível de significância, rejeita-se a hipótese nula e conclui-se que há uma correlação significativa entre as variáveis.

5. RESULTADOS

5.1 Previsões com base na Regressão Linear

- **Previsão da média da taxa de homicídios para o ano de 2019**

Para fazer a previsão da média da taxa de homicídios para o ano de 2019, calculamos a média da taxa de desmatamento em 2019 e assumimos que não tínhamos a taxa de homicídios. A média foi então atribuída à variável x e, utilizando os coeficientes da regressão linear encontrados (coeficiente angular m e coeficiente linear b), aplicamos a fórmula da regressão linear (y_previsto = m*x + b) para obter a previsão da taxa de homicídios. Em seguida, comparamos essa previsão com a média real da taxa de homicídios para o ano de 2019.

Os resultados dessa previsão foram os seguintes:

Previsão: 0.12. Valor real: 0.11

- **Previsão da taxa de homicídios para o município com maior desmatamento em 2019**

Para fazer a previsão da taxa de homicídios para o município com maior desmatamento em 2019, primeiramente, encontramos o id_municipio correspondente ao maior valor de taxa de desmatamento, que é Lago dos Rodrigues - MA. Em seguida, selecionamos a taxa de desmatamento e a taxa de homicídios para esse município e o ano de 2019. O valor da taxa de desmatamento foi atribuído à variável x, e aplicamos novamente a fórmula da regressão linear para obter a previsão da taxa de homicídios. Também comparamos essa previsão com o valor real da taxa de homicídios para o município.

Os resultados dessa previsão foram os seguintes:

Previsão: 0.13. Valor real: 0.15

- **É possível classificar uma cidade da Amazônia Legal como perigosa ou não segundo sua taxa de desmatamento?**

Essa pergunta pode ser respondida olhando para a regressão linear que fizemos para prever a taxa de homicídios com base na taxa de desmatamento. Uma vez que temos uma taxa de desmatamento x_i, obtemos, pela regressão uma taxa de homicídios estimada y_i. Com base no valor de y_i, podemos definir faixas de valores e para verificar quais valores de y_i seriam considerados perigosos ou não. Porém, é importante ressaltar que a regressão obtida explica apenas 0,07% da variância dos dados, o que não é algo tão significativo. Sendo assim, concluímos essa classificação com base em y_i, também não seria tão significativa, uma vez que a regressão também não é.

5.2 Outros Resultados

- **Quais são os pares de cidades da Amazônia Legal com taxa de desmatamento semelhante e taxa de homicídios divergentes?**

Para responder a esta pergunta, realizamos uma análise para identificar os pares de cidades da Amazônia Legal que possuem taxas de desmatamento semelhantes, mas taxas de homicídios divergentes. Primeiro, é calculada a distância euclidiana entre as taxas de desmatamento de todas as cidades. Em seguida, para cada cidade, é identificada a cidade com a taxa de desmatamento mais próxima. Depois, é calculada a diferença entre as taxas de homicídios médias dos pares de cidades. Os pares de cidades são ordenados com base nessas diferenças e os 5 pares mais divergentes são selecionados. Finalmente, os pares de cidades mais divergentes são exibidos, mostrando suas taxas de desmatamento e taxas de homicídios respectivas.

6. CONCLUSÕES

Com base nos resultados obtidos, podemos concluir que a taxa de desmatamento possui uma relação muito fraca com a taxa de homicídios na região de estudo. A baixa magnitude da correlação e o baixo valor do coeficiente de determinação sugerem que outros fatores além do desmatamento influenciam de forma mais significativa a ocorrência de homicídios na região.

Ao analisar as previsões, podemos observar que as taxas de homicídios previstas para o ano de 2019, tanto para a média geral quanto para o município com maior desmatamento, estão próximas aos valores reais. No entanto, é importante destacar que os valores previstos apresentam diferenças em relação aos valores reais.

Considerando o baixo coeficiente de determinação R^2 (0.007) obtido na regressão, juntamente com as correlações de Pearson próximas a zero entre as variáveis taxa de homicídios e taxa de desmatamento (0.08), concluímos que a regressão linear não é um modelo adequado para capturar a relação entre essas variáveis. Essa conclusão é reforçada pelos resultados das previsões, que mostram uma certa discrepância entre os valores previstos e os valores reais.

Essas previsões destacam a limitação da regressão linear ao lidar com a complexidade da relação entre a taxa de homicídios e a taxa de desmatamento. Outros fatores não considerados nesse modelo, como fatores socioeconômicos, políticos e culturais, podem ter uma influência significativa na ocorrência de homicídios, tornando a regressão linear inadequada para fazer previsões precisas nesse contexto.