Trabalho prático 2

Introdução a banco de dados

Relatório de exploração

Integrantes:

Luis Antonio Duarte Sousa

Felipe Dias De Souza Martins

João Paulo Moura Furtado

Victor Gabriel Araujo Barbosa

Marcus Vinicius Moraes Oliveira

Link para repositório : https://github.com/DuarteDvv/UFMG.ibd_tp2_data_exploration

Observações:

- *O Dicionário está no repositório em um arquivo com extensão .pdf*
- *Os metadados estão no arquivo README com extensão markdown .md*
- *Os dados brutos usados na construção estão nas pastas raw_data**
- *Scripts usados no tratamento, plotagem e conversão banco de dados estão na pasta scripts*

1. INTRODUÇÃO

O objetivo principal dessa análise é encontrar/evidenciar a correlação de dados de violência geral e violência contra jovens com dados educacionais de cada município no Brasil. A hipótese defendida é que existe uma relação entre a degradação da educação nos municípios e o aumento da violência registrada pelos órgãos federais.

2. ANÁLISE CRÌTICA DAS FONTES UTILIZADAS

Foram visitadas diversas fontes para a produção do trabalho, dentre elas, pode se citar o INEP, IBGE, IPEA e MJSP, que, apesar de serem institutos renomados no país, apresentaram vários problemas nos dados, desde a forma com que foram disponibilizados até na apuração dos mesmos.

Dito isso, alguns dos problemas encontrados foram:

- *Formato dos arquivos*

  Nenhum arquivo foi disponibilizado em formato compatível com banco de dados, a grande maioria veio em formato ODS ou em XLS, o que afeta na exportação desses dados para um sgbd e na formatação correta dos dados.

- *Cabeçário/Fonte dentro dos dados*

  Dados de educação possuíam cabeçario e fontes no final do arquivo que quebravam qualquer conversão para csv que fosse necessária. Tivemos que retirar tudo antes de converter

- *Legenda/Dicionário dos dados*

  Na maioria das vezes foi necessário usar do bom senso para saber o que alguns dados significam, nomes de colunas com abreviações, nomes de tabelas sem sentido. Um exemplo está em um dado que remete ao atraso de escolaridade pois não sabemos o que exatamente aquele dado remete ou qualquer normalização neles aplicada.

- *Dados nulos*

  Não era difícil encontrar índices com valores nulos, às vezes até em chaves estrangeiras, o que compromete o dado e faz com que a sua validação seja questionável.

- *Nulos representados como '--'*

  Nos dados de educação os itens vazios/nulos eram preenchidos com '--' e isso fez ser necessário uma substituição usando scripts python ou simplesmente o comando POSIX (tr '--' '' < data).

- *Dados faltando*

  Isso se deu em todas as tabelas de todos os dados, usando o exemplo de municípios, tinham vários municípios que não tinham dado nenhum, alguns ficaram vazios e outros não eram citados, e não era só em cidades com poucos habitantes, mas Porto Alegre, capital do Rio Grande do Sul e uma das maiores cidades do país tinham dados incompletos sobre a educação.

- *Falta de critérios expostos*

  Aliado com os dados que não tinham legendas, alguns dados não mostravam o seu critério, como por exemplo, homicídio de jovens, não se sabe ao certo qual a definição de jovem, se é contabilizado crianças, se pessoas com 23 anos são consideradas jovens, entre outras dúvidas que foram deixadas.

- *Redundância*

  Para a execução do trabalho foi necessário excluir diversas redundâncias, várias colunas que não diziam nada estavam presentes, como por exemplo, uma tabela que era só sobre 2021 tinha a coluna ano, onde todos eram "2021".

- *Problemas na tipagem dos dados*

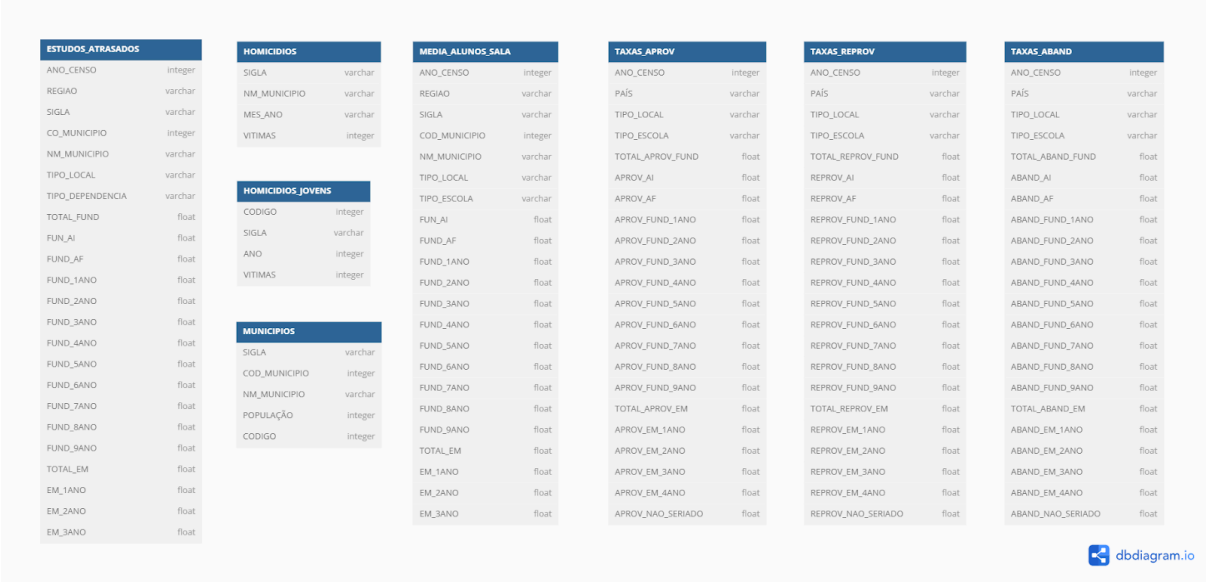
  Foi necessário diversos casting internos junto com criações de novas tabelas para conversão dos dados para o tipo de fato analisável e isso se deve provavelmente em relação aos pontos e vírgulas que surgiam com a conversão de XLS. Por exemplo, as colunas de taxas eram todas consideradas como string e foi necessário casting com novas tabelas.

- *Dados fracionados*

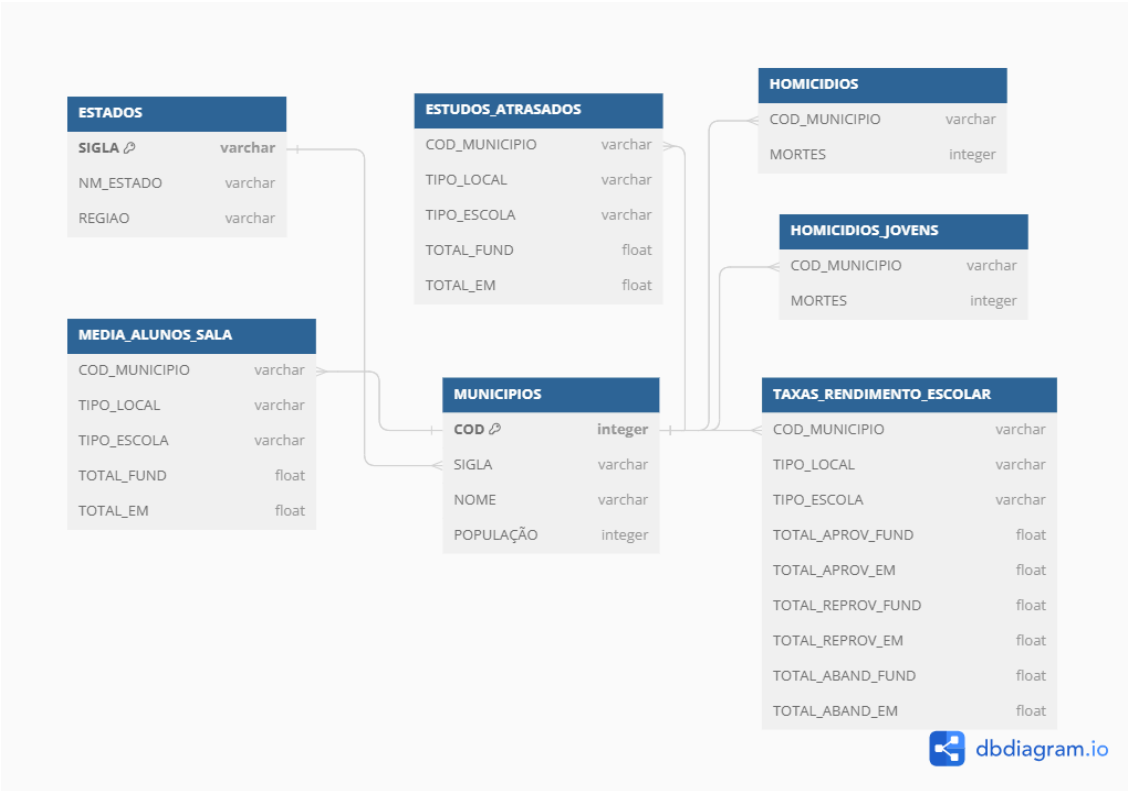
  Isso aconteceu principalmente na tabela municípios, onde tinham alguns dados que estavam separados, como por exemplo o código do município e o código do estado, eles ficavam separados, e o problema disso é que, para servir como chave primária e fazer sentido com as outras tabelas (educação, homicídios), o código deveria aparecer junto. Dito isso, tivemos que criar uma nova coluna que seria a concatenação das duas colunas de strings e depois apagar as antigas.

3. BANCO DE DADOS E EXPLORAÇÃO

- Esquema antes do tratamento dos dados:



- Esquema após o tratamento dos dados:



O objetivo desse projeto foi explorar os dados públicos, mostrando a correlação entre a violência e a educação no Brasil.

Para tal meta, foram resgatados e tratados dados referentes ao assunto, como informações sobre porcentagem de alunos com estudos atrasados, a média de alunos em cada ano de ensino e dados sobre a porcentagem de aprovação, reprovação e abandono em cada município do país. Além disso, foram buscados dados de violência, nesse caso o número de homicídios total, e a quantidade desses homicídios que foram cometidos por jovens, na faixa etária de 15 a 29 anos.

A análise integrada dos dados revelou algumas correlações significativas entre a violência e a educação no Brasil. Observou-se que municípios com maior porcentagem de alunos com estudos atrasados tendem a apresentar maiores taxas de homicídios, proporcionalmente com sua população, especialmente aqueles cometidos por jovens.

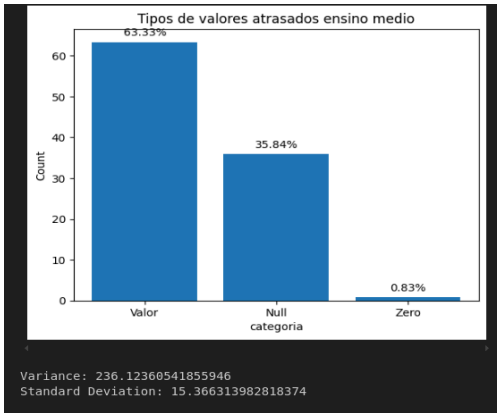
Outro dado relevante foi a relação entre as taxas de aprovação, reprovação e abandono escolar com os índices de violência. Municípios com altas taxas de abandono escolar frequentemente apresentaram maiores índices de homicídios juvenis. Da mesma forma, a reprovação escolar mostrou uma correlação com a violência, indicando que alunos que repetem de ano podem estar mais suscetíveis a se envolverem em situações de risco que levam à violência.

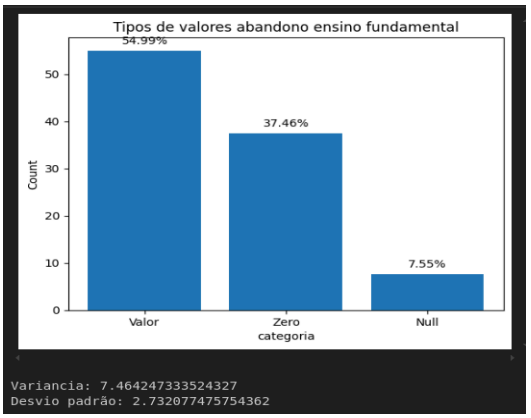
Isso sugere que o sucesso acadêmico pode ser um fator protetivo contra a violência, possivelmente devido ao aumento de oportunidades e ao engajamento escolar positivo que mantém os jovens fora de situações de risco.

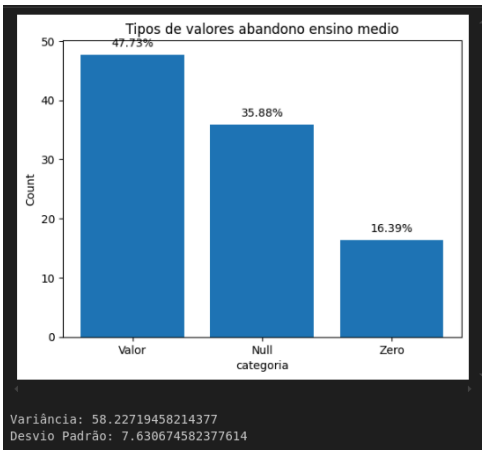
Em resumo, a análise integrada dos dados indica que a educação e a violência estão fortemente correlacionadas no Brasil. Atrasos escolares, reprovação e abandono estão associados a maiores índices de violência. Esses achados ressaltam a importância de políticas públicas focadas na melhoria da educação como uma estratégia para a prevenção da violência.

Em primeiro lugar é importante dizer que os dados que serão cruzados serão entre
taxa_estudos atrasados x homicidios
taxa_aprovacao x homicidios
taxa_reprovacao x homicídios
taxa_abandono x homicidios

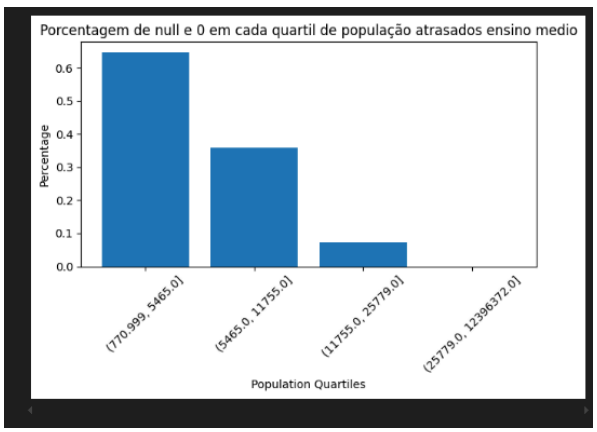
A primeira análise feita feita nos dados era a sua distribuição em três tipo de valores NULL, 0 e o resto dos valores

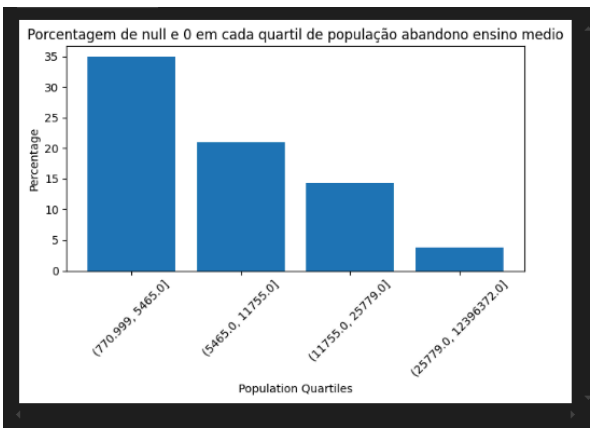






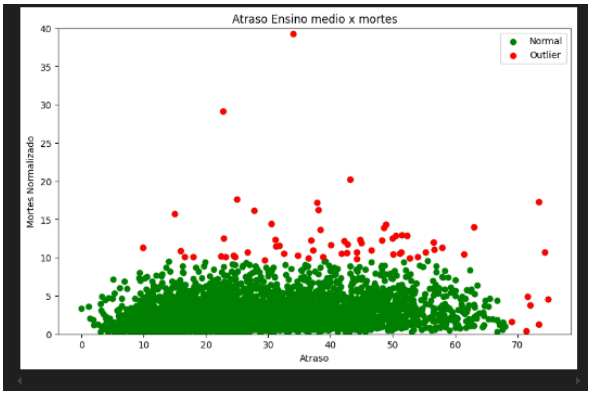
Avaliando os gráficos e seus valores é preciso dizer que como grande parte dos valores são NULL (> 30%) a análise feita abaixo não é exata, já que existem muitos valores faltantes dentro de nosso dados, e pode indicar que a coleta de dados dos meios públicos brasileiros não são eficientes, com a hipótese que os valores NULLS ou zero possam indicar ausência de dados e eles estão principalmente em cidades do interior pouco populosas, foram criados mais duas tabelas de barras, uma para atrasados e outra abandono, dividindo a população em quatro quartis e fazendo a porcentagem de cada quartil de valores NULL mais valores 0

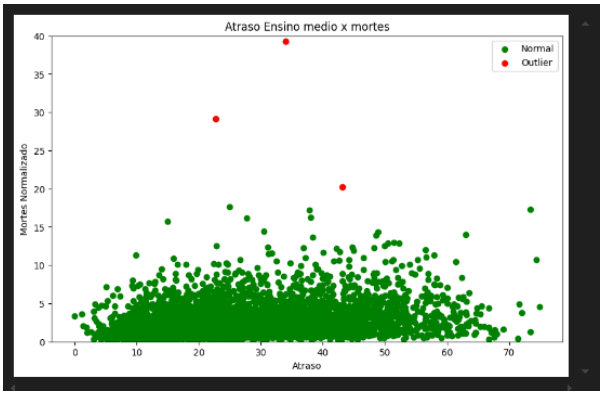




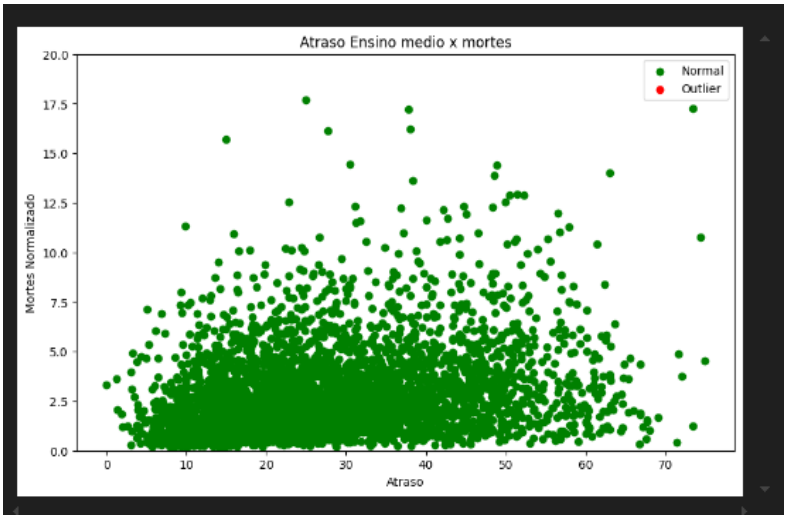
Nesses gráficos provam a correlação que quanto menor a cidade mais chance de as informações não serem coletadas, então talvez a análise não tenha dados exatos devido essa falha no processo de coleta da informação e falta de dados para municípios pequenos.

Após essa análise inicial foi preciso identificar valores discrepantes (outliers) em primeiro lugar foram feitos gráficos de dispersão entre os valores mencionados no começo da análise exploratória



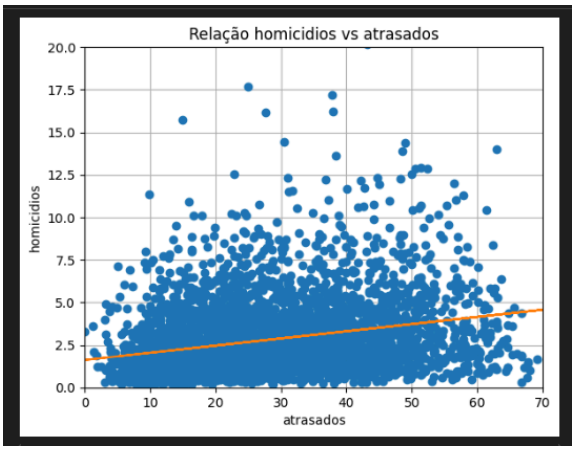


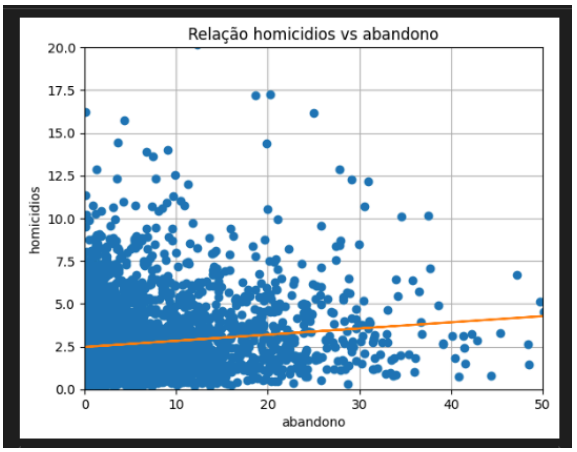
Em primeiro lugar é preciso dizer que as mortes estão normalizadas nesse gráfico, com isso em mente foi calculado o zscore dos dois eixos para calcular os outliers no primeiro gráfico foi usado o valor de 3 desvios padrão para calcular os outliers, como acredito que os dados estão muito concentrados na parte de baixo do gráfico, muitos dados que seriam interessantes para análise estão sendo classificados como outliers, para que esses dados fossem utilizados mudei o valor para o cálculo para 6 desvios padrões, que pegou apenas os mais extremos, resultando em

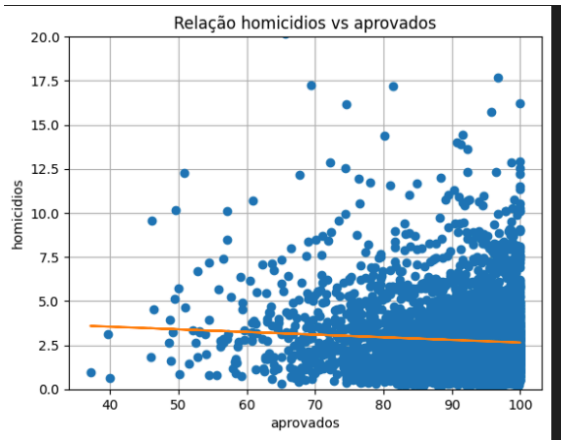


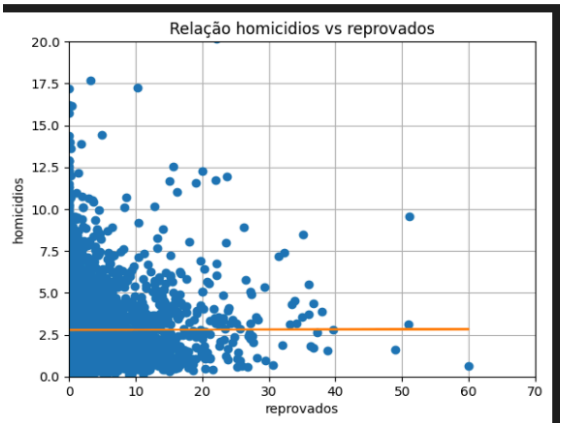
que é um gráfico bem mais compacto e que os outliers não vão acarretar em significativas mudanças, nos outros gráficos foram aplicados a mesma técnica.

Em seguida, para ver se existe correlação entre esses indicadores e os homicídios, foi feito nos gráficos mostrados anteriormente o processo de regressão linear.



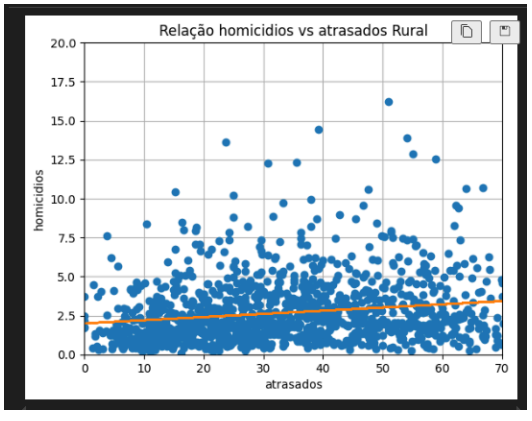


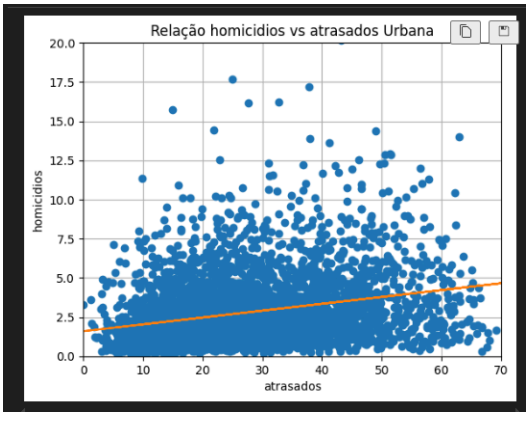


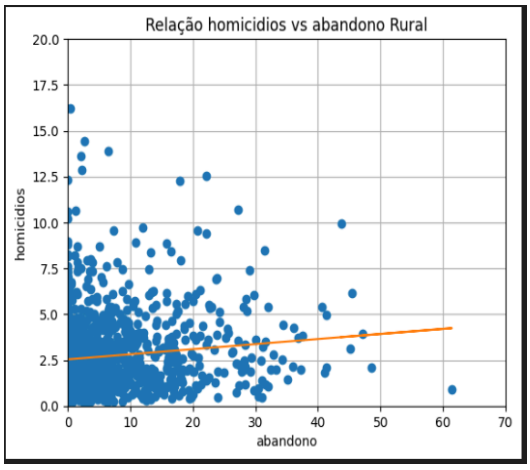


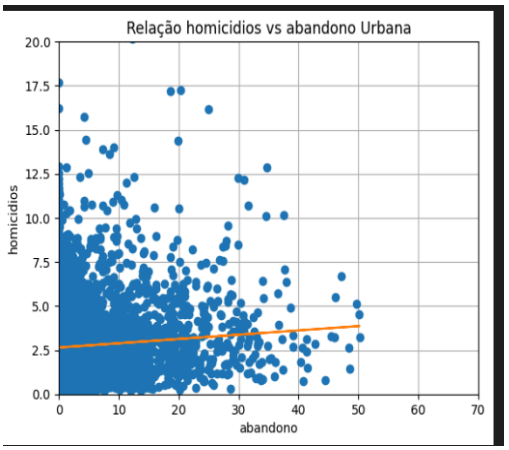
Como mostrado nos gráficos realmente existe uma correlação entre a quantidade de homicídios e os índices, apesar da linha não crescer tanto é perceptível que ela tem uma taxa de crescimento, o que mostra a correlação entre os dois, apenas em questão do gráfico de reprovados que a linha cresce com um angulação muito baixa parecendo ser uma linha reta.

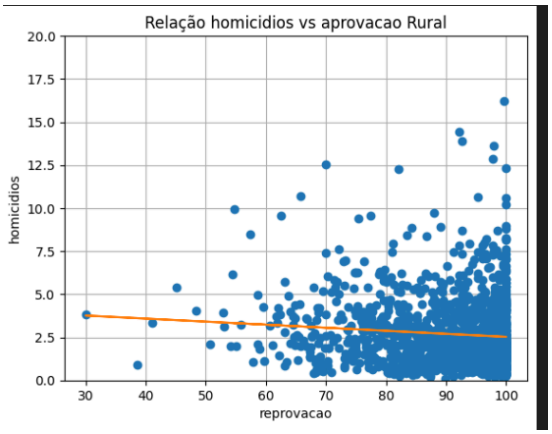
Além disso foram feitos gráficos para os indicadores em municípios rurais e urbanos

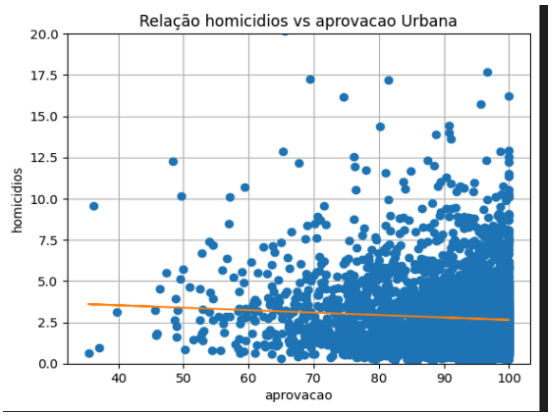




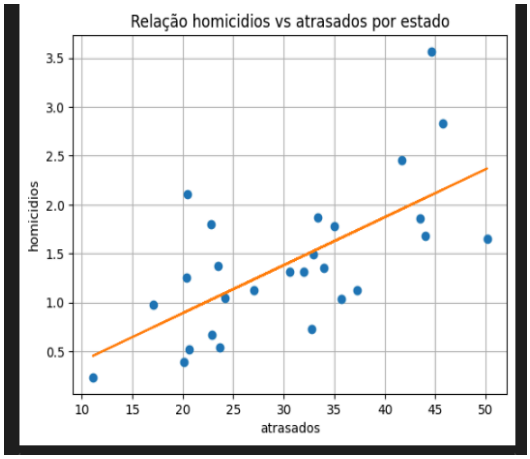


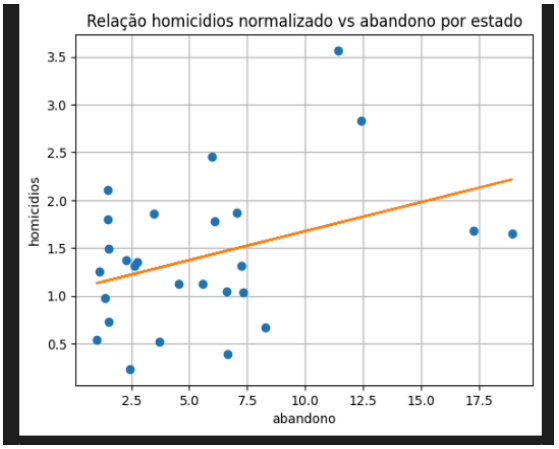


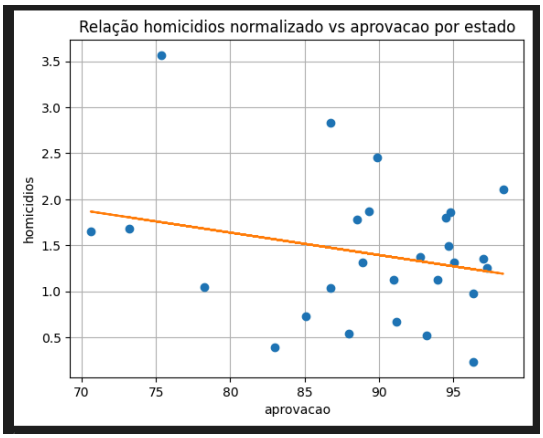


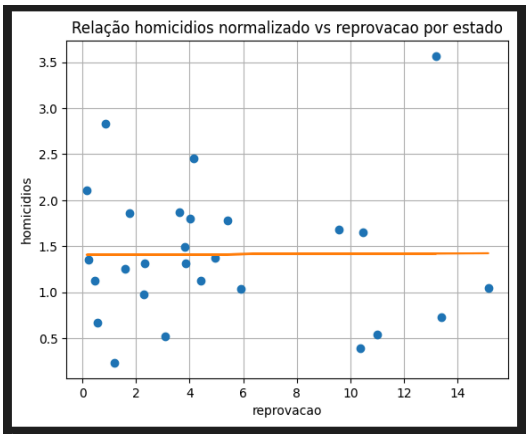


É possível perceber que a correlação em áreas urbanas é bem maior que em municípios rurais, mostrando que a violência é bem mais encontrada em áreas urbanas dos municípios. Por fim, foi criado um gráfico para correlações agrupado pelos estados, assim tirando a concentração de muitos dados e deixando apenas 27, com isso é possível ver mais claramente as estatísticas.









É possível dizer que as regressões lineares são mais claras quando os dados são agrupados por estado

Conclui-se com as análises que existe correlação entre a falta de estudo e a quantidade de violência dentro do contexto brasileiro, mas além disso é preciso perceber que pela grande quantidade de valores 0 e nulo que a coleta de dados feita pelos órgãos brasileiros é falha, assim deixando buracos nos dados e consequentemente deixando sua análise mais difícil e menos precisa.