# Atividade 02 - Análise Exploratória

#### Tema

Análise da criminalidade em Curitiba - Evolução temporal, distribuição espacial e correlação com os indicadores de renda, segurança e desenvolvimento das regionais da cidade de Curitiba.

# Equipe

Nome: CSI Curitibanos

Repositório GitLab da equipe: https://gitlab.com/thiagorbernardo/IntroCD

Membros:

<ul> <li>João Felipe Sarggin Machado,</li> </ul>	1905023,	jfsmachado,
joaomachado@alunos.utfpr.edu.br,	EC,	UTFPR
<ul> <li>João Lucas Mizuguchi Ferreira,</li> </ul>	2135965,	jlmizug,
joaolucasferreira@alunos.utfpr.edu.br	, EC,	UTFPR
<ul> <li>Pedro Sodré dos Santos,</li> </ul>	2126745,	pedro_sodre,
psantos.2000@alunos.utfpr.edu.br,	EC,	UTFPR
<ul> <li>Thiago Ramos Bernardo,</li> </ul>	2126753,	thiagorbernardo,
thiagobernardo@alunos.utfpr.edu.br.	EC.	UTFPR

#### Obtenção e Processamento de Dados

Primeiramente, o foco seria a análise da relação entre os parâmetros/índices dos bairros de Curitiba, com os crimes. Para isso, foram pegos estes dados através do site <a href="http://infocuritiba.ippuc.org.br/">http://infocuritiba.ippuc.org.br/</a>, porém, existem muitos dados que estão indisponíveis, então métricas e índices como quantidade de empregos, PIB, PIB per capita, IDH, IFDM (Índice Firjam de Desenvolvimento Municipal, verifica índices socioeconômicos dos municípios brasileiros), entre outros apareceram como a string: "..." (vazio) no dataset, além do dataset estar com os parâmetros nas linhas ao invés das colunas, então foi feita uma transposição (transformar as linhas em colunas, e colunas em linhas), as colunas com valores nulos ("...") foram excluídas, e também, a coluna "ANO", que, como só havia dados de 2010 disponíveis, era uma coluna com único valor, e por isso foi excluída também, e vamos induzir que não houve grandes mudanças nestes parâmetros na última década. Dessa forma o dataset sofreu uma transformação como se pode ver na imagem abaixo:

	Variável	Ano	Abranches	Água Verde	Ahú	Alto Boqueirão	Alto da Glória	Alto da Rua XV	Atuba	Augusta	 Sítio Cercado	Taboão	Tarumã	Tatuquara	Tingui	Uberaba	Umbará	Vila Izabel	Vista Alegre	Xaxim
0	População Censitária - Total (Habitante)	2010.0	13.189	51.425	11.506	53.671	5.548		15.935	6.598		3.396	8.072	52.780	12.319	72.056	18.730	11.610	11.199	57.182
1	Classe de Rendimento - Pessoas de 10 Anos ou M	2010.0		12.044	2.920	14.447		1.936	4.097	1.798	31.281			14.962	3.065	19.815	5.421	2.876	2.947	15.940
2	Rendimento Nominal Mensal Médio das Pessoas de			3.170,07	3.054,92	867,36	3.054,18	2.854,97	1.203,26	731,64	726,28	1.381,97	2.100,23	590,06	1.370,39	1.101,82	762,65	2.467,22	2.036,11	1.040,56
3	Rendimento Nominal Mensal Mediano das Pessoas	2010.0	650,00	1.511,00	1.500,00	650,00	2.000,00	1.500,00	700,00	520,00	600,00	700,00	1.000,00	510,00	800,00	600,00	600,00	1.500,00	810,00	700,00
4	Empregos - Total (Habitante)	2010.0																		
5	PIB (R\$1.000,00) (Real)	2010.0																		
6	PIB - Per Capita (R\$1,00) (Real)	2010.0																		
7	Alfabetização - Taxa de Analfabetismo - Total	2010.0		0,82	0,87		0,84	0,99		4,85	4,44									3,07
8	IDHM - Geral (Unidade)	2010.0																		
9	IDHM - Renda (Unidade)	2010.0																		
10	IDHM - Longevidade (Unidade)	2010.0																		
11	IDHM - Educação (Unidade)	2010.0																		
12	IFDM - Geral (Unidade)	2010.0																		
13	IFDM - Educação (Unidade)	2010.0																		
14	IFDM - Saúde (Unidade)	2010.0	-	-	-		-	_		-	 -	-			-	-		-	-	

Figura 1. Raw Dataset (antes da transformação)

	 		Rendimento Nominal Mensal Mediano das Pessoas de 10 Anos ou Mais de Idade - Total (Real)	
				Artabetização - taxa de Anartabetismo - total (Percentual)
ABRANCHES				3,71
ÁGUA VERDE				0,82
AHÚ		3.054,92	1.500,00	0,87
ALTO BOQUEIRÃO				3,36
ALTO DA GLÓRIA			2.000,00	0,84
_				_
UBERABA				3,51
UMBARÁ			600,00	4,37
VILA IZABEL			1.500,00	1,15
VISTA ALEGRE				1,87
XAXIM		1.040,56		3,07

Figura 2. Dataset limpo

Também foi feita a mudança dos nomes dos bairros para *uppercase*, para posteriormente fazer a junção das duas tabelas.

Outra tabela também adquirida e utilizada, foi a de ocorrências em Curitiba, disponibilizados pela própria Guarda Municipal da cidade, disponibilizada no site <a href="https://www.curitiba.pr.gov.br/dadosabertos/busca/?termo=guarda">https://www.curitiba.pr.gov.br/dadosabertos/busca/?termo=guarda</a>, também foi utilizada uma tabela semelhante com a geolocalização (Latitude e Longitude) dos crimes, de um dataset disponibilizado pelo professor.

Visualizando as coisas básicas do dataset, foi visto que existiam muitas (185) naturezas de crime (Roubo, Furto, Desacato, Invasão, etc.), então, foi considerado as naturezas mais relevantes e comuns, reduzindo o número de naturezas de 185 para 41, para serem analisadas.

Após essas considerações, as análises foram divididas em dois caminhos: de um lado, teve uma busca para entender a relação entre os índices dos bairros com os tipos de crimes realizados, e para isso, foi feito uma junção do dataset dos índices dos bairros com os crimes através da coluna que continha o nome do bairro; do outro lado, foi feito uma análise buscando analisar a relação entre dias e horários, com os crimes realizados, dessa forma a análise foi dividida em 2 jupyter notebooks.

#### Cobertura e distribuição dos dados

Para verificar a distribuição dos dados, antes de tudo foi verificado os crimes mais cometidos dentro da cidade, verificando o top 10 crimes mais cometidos em Curitiba, que pode ser visto pela tabela abaixo:

NATUREZA1_DESCRICAO	TOTAL_POR_CRIME
Roubo	17925
Furto	8346
Ameaça	7762
Agressão fisica/verbal	6573
Violação de Medida Protetiva Lei Maria da Penha	3681
Risco de acidente / à vida	2755
Alagamento	2396
Atos obscenos/libidinosos	1800
Disparo de arma	1358
Lesão Corporal	936

Figura 3. Tabela de crimes mais cometidos

Dessa forma, é verificado que a maior parte de crimes cometidos na cidade são de roubos e furtos, além de uma grande quantidade de ameaças e agressões.

Para entender melhor, e termos gráficos mais relevantes foi feito um agrupamento dos tipos de ocorrências, juntando roubo e furto, crimes de pudor etc, como é possível ver na tabela abaixo:

	NATUREZA	TOTAL_POR_CRIME
7	ROUBO	26839
0	AGRESSAO	15361
6	RISCO_VIDA	4382
1	CRIME_MULHER	3681
5	PUDOR	2805
8	SEQUESTRO	415
4	MORTES	141
2	ESTUPRO	108
3	EXPLORACAO	17

Figura 4. Tabela de crimes mais cometidos

Após isso, como números totais não são bons de fazer a análise, foi feita a proporção de crimes com sua população, adicionando uma nova coluna ('Proporcao') no dataset. Como é possível ver na tabela abaixo.

	ATENDIMENTO_BAIRRO_NOME	POPULACAO_TOTAL	TOTAL_POR_BAIRRO	PROPORCAO
37	RIVIERA	289	2	0.006920
30	LAMENHA PEQUENA	1056	10	0.009470
18	CASCATINHA	2161	36	0.016659
19	CAXIMBA	2522	74	0.029342
28	HUGO LANGE	3392	51	0.015035
10	BOM RETIRO	5156	93	0.018037
29	JARDIM SOCIAL	5698	47	0.008249
36	PRADO VELHO	6077	504	0.082936
3	AUGUSTA	6598	192	0.029100
15	CAMPINA DO SIQUEIRA	7326	542	0.073983
32	ORLEANS	8105	110	0.013572
23	FANNY	8415	236	0.028045
1	ALTO DA RUA XV	8531	315	0.036924
13	CACHOEIRA	9314	128	0.013743
7	BATEL	10878	481	0.044218
25	GANCHINHO	11178	262	0.023439
43	VISTA ALEGRE	11199	169	0.015091
26	GUABIROTUBA	11461	228	0.019894
33	PAROLIN	11554	589	0.050978
42	VILA IZABEL	11610	145	0.012489
40	TINGUI	12319	195	0.015829
11	BUTIATUVINHA	12876	160	0 012426

Figura 5. Tabela de Bairros com a população, quantidade total de crimes, e a proporção (Crimes/População)

Também foi verificada a distribuição do tipo de crime com maior quantidade que é o roubo, juntamente com um dos parâmetros considerados primariamente importantes na análise, que é o rendimento mensal mediano, o que significa que pelo menos metade da população recebe menos que esse valor. Dessa forma foi-se obtido a tabela abaixo:

	ATENDIMENTO_BAIRRO_NOME	POPULACAO_TOTAL	RENDIMENTO_MENSAL_MEDIANO	TOTAL_ROUBOS	PROPORCAO
938	TATUQUARA	52780	510.0	117	0.002217
478	CAXIMBA	2522	510.0	6	0.002379
874	PRADO VELHO	6077	510.0	243	0.039987
643	GANCHINHO	11178	510.0	18	0.001610
93	AUGUSTA	6598	520.0	23	0.003486
446	CAMPO DE SANTANA	26657	580.0	55	0.002063
989	UBERABA	72056	600.0	122	0.001693
290	BUTIATUVINHA	12876	600.0	30	0.002330
726	LAMENHA PEQUENA	1056	600.0	3	0.002841
794	PAROLIN	11554	600.0	316	0.027350
849	PINHEIRINHO	50401	600.0	524	0.010397
550	CIDADE INDUSTRIAL DE CURITIBA	172822	600.0	623	0.003605
366	CAJURU	96200	600.0	849	0.008825
331	CACHOEIRA	9314	600.0	13	0.001396
21	ABRANCHES	13189	650.0	46	0.003488
769	ORLEANS	8105	700.0	22	0.002714
616	FAZENDINHA	28074	700.0	217	0.007730
70	ATUBA	15935	700.0	44	0.002761
819	PILARZINHO	28480	700.0	46	0.001615
1052	XAXIM	57182	700 0	582	0 010178

Figura 6. Tabela de Bairros com a população, rendimento mensal mediano, total de roubos e a proporção (Roubo/População)

### Identificação de Padrões Iniciais

Foi tentado verificar se existia alguma relação entre os dias da semana, se haveria uma maior ocorrência de crimes em algum dia específico, dessa forma, os crimes foram agrupados pelas quantidades de dias da semana, gerando o seguinte gráfico:

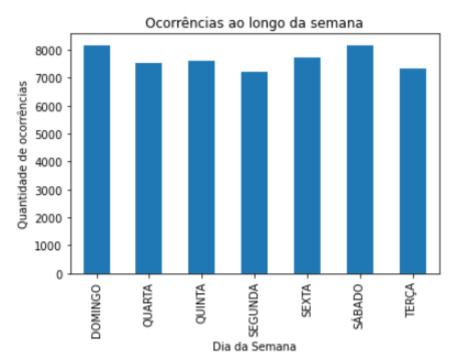


Figura 7. Gráfico de Barras mostrando a ocorrências crimes por dia da semana

Com esse gráfico, foi possível visualizar que não há uma relação tão explícita de um período da semana que ocorre mais crimes que outros, é possível verificar que no início da semana (Segunda e Terça) são os dias que possuem menos ocorrências. Porém a diferença não chega a ser nem de 10% em relação aos outros dias.

Também foi verificado se, da mesma forma que foi feita com os dias da semana, foi feita com os horários do dia, buscando em que período do dia seria mais provável de ocorrer algum tipo de crime, no entanto, para isso, foi necessário separar os horários dos minutos, montando uma nova coluna com apenas a hora registrada para fazer a análise. E, após isso, foi verificado os seguinte gráficos de horas do dia por quantidade de crimes cometidos:



Figura 8. Gráfico de Barras mostrando a ocorrência de crimes em cada hora

Dessa forma, como já era de se esperar pelo senso comum, a maior quantidade de crimes ocorrem no período da noite, e há uma grande diferença do período da madrugada juntamente com início da manhã (da 00h às 6h), com o período da noite (das 19h às 23h), com mais 3,5 vezes mais crimes ocorrendo no período noturno. Para entender melhor essas ocorrências foi separado pelo tipo e plotado a quantidade de cada um.

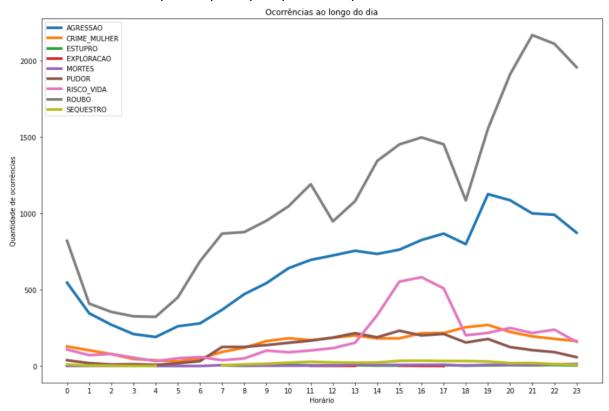


Figura 9. Gráfico de linha mostrando a ocorrência de cada crime em cada hora

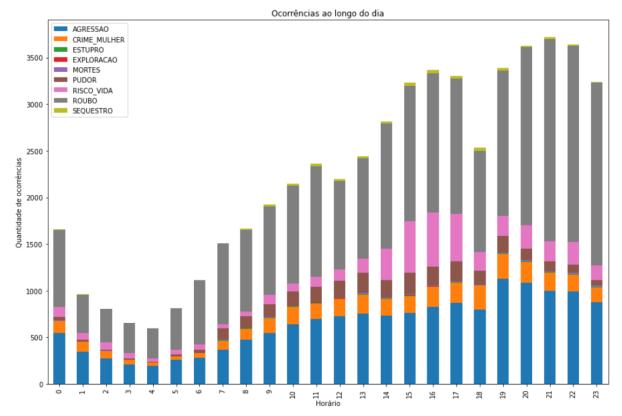


Figura 10. Gráfico de barra empilhado mostrando a ocorrência de cada crime em cada hora

É possível ver uma maior quantidade de roubos, sendo o tipo de reclamação/crime mais comum independente do horário. Roubos aumentam bastante depois das 10h e tem um pico às 21h, já agressão tem um pico às 19h. Uma observação interessante é que ocorrências do tipo de risco a vida ocorrem mais entre as 15h e 17h. Porém, crimes a mulher e outras categorias são constantes durante o dia, apenas diminuindo na madrugada

Depois partindo para uma análise dos bairros para os crimes foi feito uma agregação dos bairros que gerou o seguinte gráfico:

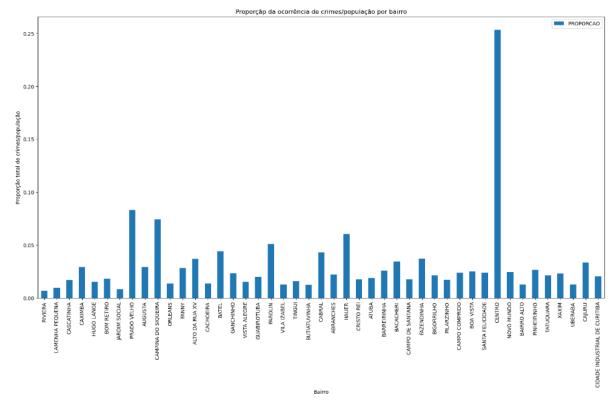


Figura 9. Gráfico de Barras mostrando a quantidade de crimes em cada bairro.

Como é possível observar no gráfico, o bairro com a maior quantidade proporcional de crimes é o Centro, que é o nono bairro mais populoso, já os outros 8 bairros mais populosos possuem uma quantidade menor de crimes. Com isso é possível concluir que o tamanho da população é um fator que não tem um peso grande na ocorrência de crimes. Também foi possível verificar que a área mais afetada é o centro, com cerca de 0,25 crimes por habitante, maior que o dobro da segunda maior, o que significa uma grande taxa de crimes (podemos aproximar que a cada 4 habitantes no centro, 1 foi vítima de crime).

Depois foi feita uma análise do crime mais cometido na cidade: o roubo. Como é possível observar no gráfico, o bairro com a maior quantidade de roubos é o Centro, que é o nono bairro com maior renda mensal mediano, já os outros 8 bairros com maior renda mensal mediano possuem uma quantidade menor de roubos. Com isso é possível concluir que a renda mensal mediana é um fator que não tem um peso grande na ocorrência de roubos. Também foi possível verificar que a área mais afetada é o centro, com cerca de 0,10 roubos por habitante, maior que o dobro da segunda maior, o que significa uma grande taxa de roubos (podemos aproximar que a cada 10 habitantes no centro, 1 foi vítima de roubo).

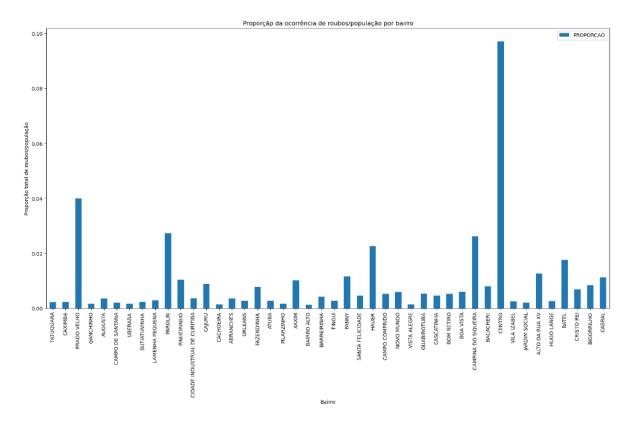
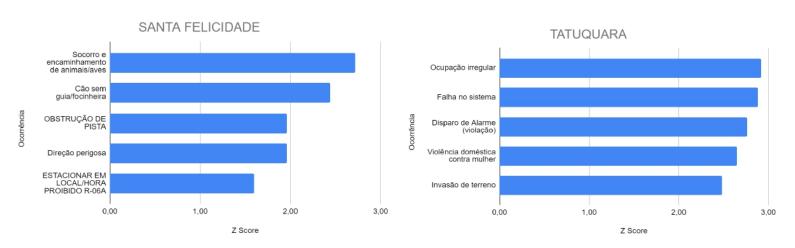
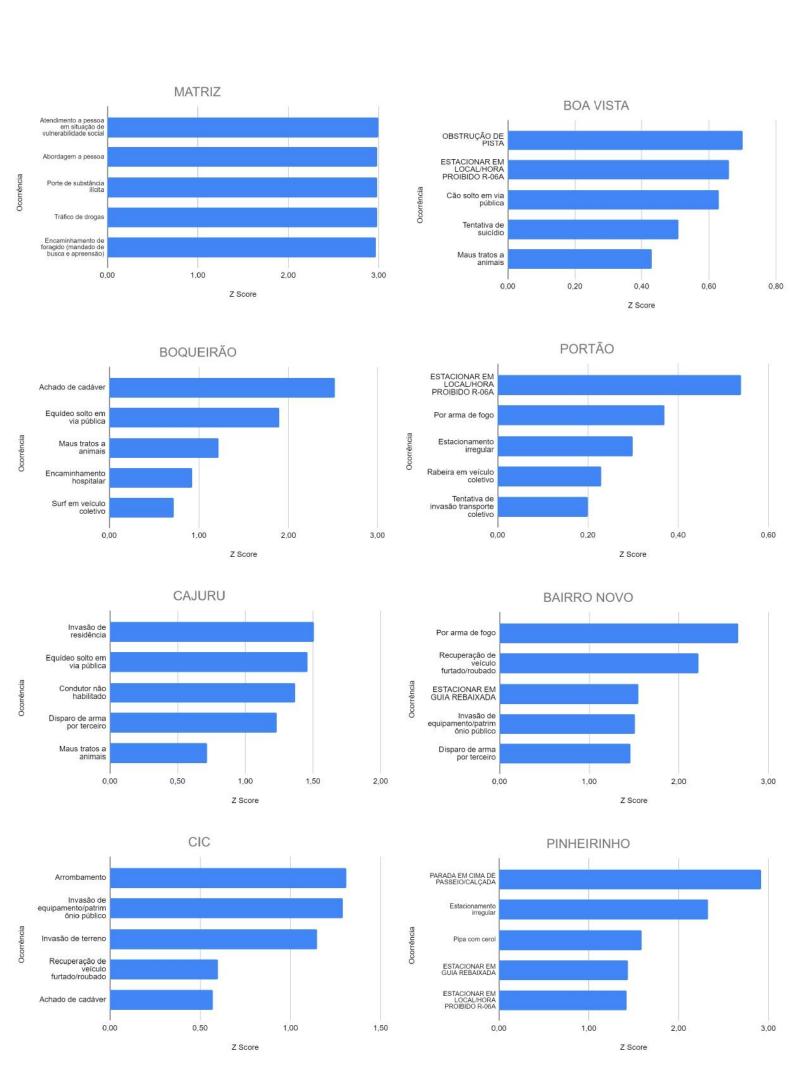


Figura 10. Gráfico de Barras mostrando a quantidade de roubos em cada bairro.

Em seguida foi contabilizado quantas ocorrências cada tipo teve em cada regional e normalizá-las para cada 10 mil habitantes que moram naquela regional. Com esse valor, calculou-se a porcentagem em que cada regional contribui para cada tipo de ocorrência e então calculou-se o desvio padrão dessa porcentagem entre as regionais.

Em um caso onde os tipos de ocorrência são perfeitamente distribuídos, esperaria-se que cada regional contribuisse com 10% da quantidade normalizada de casos para cada 10 mill habitantes, mas este não foi o caso, sendo possível notar então, que em algumas regionais possuem certas particularidades que mais a diferenciam das demais, segue os gráficos dos Z-Scores obtidos nessa análise para cada regional:





Interessante notar como as regionais concentram tipos diferentes de ocorrências:

Santa Felicidade: Animais e direção imprópria de veículos.

**Tatuquara**: Terrenos, alarmes e violência doméstica contra mulheres.

Matriz: Drogas, atividades suspeitas e ajuda a moradores de rua.

Boa vista: Veículos, animais e tentativas de suicídio.

Boqueirão: Cadáver, animais, encaminhamento hospitalar e surf em ônibus.

Portão: Estacionamento de carros, porte de arma de fogo e relacionados aos ônibus.

Cajuru: Invasão de residência, animais e disparo de arma de fogo.

**Bairro Novo**: Arma de fogo, veículos roubados e invasão de patrimônio público. **CIC**: Arrombamentos, invasões, recuperação de veículos furtados e cadáveres.

Pinheirinho: Estacionamento de carros e pipas com cerol.

#### Perguntas de pesquisa e explorações iniciais

Existe uma relação entre a quantidade de crimes e os dias da semana? Com base no descrito na identificação de padrões iniciais é possível ver que não.

Existem horários mais propícios para os crimes? É possível ver que sim, já que existe uma concentração bem perceptível dos crimes no período noturno, e uma queda significativa no período da madrugada e manhã.

Qual o motivo das ocorrências de risco à vida crescerem bastante entre às 15h e 18h? Ainda não foi possível identificar.

Existe um horário propício para denúncias de crimes à mulher? Aparentemente não, apenas não ocorre muito no horário da madrugada.

É possível verificar uma relação entre população e renda e quantidade de crimes? Com base no que foi visto, é possível verificar que não há uma relação explícita, por não possuir um padrão específico.

## Discussão e próximos passos

Os dados de **crimes vs pessoa** reportados são parcialmente completos, foi possível entender que talvez não seja uma tarefa fácil relacionar os tipos de crimes com sexo, idade, etc. Porém vai ser possível criarmos mapas com as ocorrências e entender as relações com os bairros.

Também é possível verificar que é possível fazer uma grande quantidade de análises em relação aos tipos de crimes cometidos (qual o crime mais cometido em cada ajudbairro, se existe alguma relação entre os índices do bairro e o seu crime mais cometido, etc).

Verifica-se que também pode haver uma relação grande entre a quantidade proporcional de crimes em um determinado bairro, e sua distância em relação ao centro, isto é, verificando que os índices dos bairros não são muito afetados pela quantidade proporcional de crimes, pode ser que haja uma relação entre a distância.

Entendemos que não será necessário mudar o tema, foi adequado e essa análise exploratória nos deu muito insumo para vermos os pontos que podemos explorar. Será

interessante talvez buscar mais dados geográficos, dados da cidade de Curitiba (se possível).