Alunos: Carla Beatriz Ferreira, Gabriele Pinheiro Sá, João Marcos Ribeiro Tolentino, Manuela Monteiro Fernandes de Oliveira e Vitor Terra Mattos do Patrocínio Veloso
Disciplina: Introdução a Banco de Dados (TZ)
Professor: Clodoveu Augusto Davis Junior
Data: 30/11/2023

TRABALHO PRÁTICO 2 - RELATÓRIO

0. INTRODUÇÃO

Este trabalho tem como objetivo a análise e a exploração de conjuntos de dados públicos, proporcionados por meio da Lei de Acesso à Informação (Lei nº 12527/2011, ou LAI), que garante a transparência das informações governamentais. Dessa forma, foram escolhidos dois conjuntos de dados, sendo um sobre focos de calor e outro sobre população – tendo, ambos, abrangência nacional, sendo produzidos por órgãos de pesquisa brasileiros e referentes a um período de tempo de aproximadamente 20 anos: o primeiro, de 1998 a 2017, e o segundo, de 1991 a 2021. Além disso, foi criado um dataset Estado, para relacionar os dois bancos citado, utilizando dados dos estados brasileiros, incluindo o Distrito Federal. Usamos esses dados para entender a relação entre a quantidade populacional dos estados no Brasil com os focos de calor distribuídos ao longo do país.

Este relatório divide-se nos seguintes módulos:

1. Preparação dos dados
2. Definição dos objetivos
3. Análise descritiva
4. Identificação de valores discrepantes
5. Análise de correlação
6. Conclusões
7. Bibliografia e bancos utilizados

1. PREPARAÇÃO DOS DADOS

Com o objetivo de garantir que os dados de ambos os bancos pudessem ser analisados e relacionados corretamente, houve uma padronização dos atributos de algumas entidades. Por exemplo, a quantidade de focos de calor era lida como um número decimal por conta do uso do ponto para separar as ordens, de forma que foi necessário retirá-lo para garantir uma computação correta das informações. Além disso, o nome dos estados no *dataset* relativo a focos de calor foi substituído pela sua sigla, para possibilitar a conexão com o banco referente à população, que se referia aos estados por suas siglas. Os nomes de estados e meses que continham acentos e/ou caracteres especiais foram normalizados em ambos os bancos de

dados, com o objetivo de facilitar as consultas. Ademais, houve a remoção de linhas duplicadas presentes no banco de dados de focos de calor.

Abaixo, nas figuras 1 a 4, há a descrição de alterações feitas na tabulação, remoção de aspas e duplicatas, e adição de parênteses nos respectivos arquivos mencionados.

Figura 1 - Substituição da tabulação para vírgula no arquivo rfincendiosflorestaisfocoscalorestados1998-2017.csv.

Figura 2 - Remoção das aspas no arquivo rfincendiosflorestaisfocoscalorestados1998-2017.csv.

```python
dados = '''Ano, Estado, Mes, Numero, Periodo
1998, Acre, Janeiro, 0, 01/01/1998
1999, Acre, Janeiro, 0, 01/01/1999
2000, Acre, Janeiro, 0, 01/01/2000
2001, Acre, Janeiro, 0, 01/01/2001
2002, Acre, Janeiro, 0, 01/01/2002
(...)
'''

# Separando as linhas
linhas = dados.split('\n')

# Adicionando parênteses ao início e final de cada linha
linhas_formatadas = ['(' + linha.strip() + ')' for linha in linhas]

# Juntando as linhas formatadas de volta
dados_formatados = '\n'.join(linhas_formatadas)

# Exibindo os dados formatados
print(dados_formatados)
```

Figura 3 - Código usando Python para adição de parênteses nos dados .csv para importação no banco de dados em SQL manualmente, para evitar o uso de um SGBD como PostgreSQL.

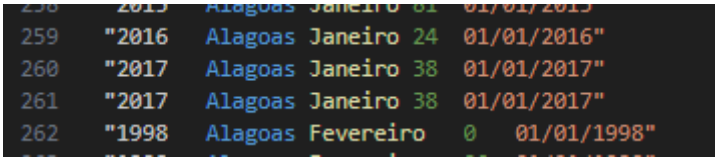

Figura 4 - Remoção da linha de dados duplicado para Alagoas em Janeiro de 2017, percebido depois da distinção de (ano, estado, mes) como chave.

Além disso, para a preparação dos dados, foi feita a recuperação de um esquema conceitual dos dados obtidos, por meio do software MySQL, a fim de permitir melhor direcionamento na construção das novas tabelas a partir dos *datasets* selecionados. A figura abaixo ilustra o diagrama gerado, apresentando 3 tabelas, sendo elas: Estado, FocosDeCalor e PopulacaoDosEstados, com seus respectivos atributos. A chave amarela indica que o atributo é uma Primary Key e o losango laranja indica que é uma Foreign Key.

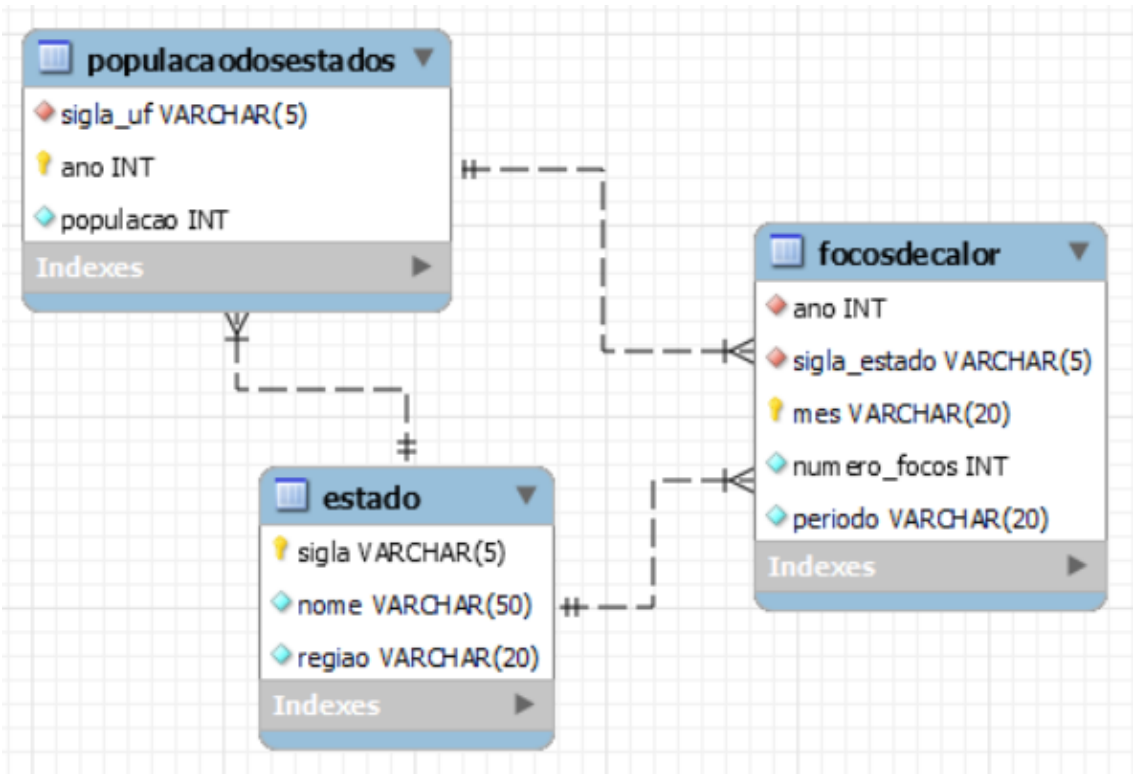

Figura 5 - Diagrama relacional do banco de dados

Alunos: Carla Beatriz Ferreira, Gabriele Pinheiro Sá, João Marcos Ribeiro Tolentino, Manuela Monteiro Fernandes de Oliveira e Vitor Terra Mattos do Patrocínio Veloso
Disciplina: Introdução a Banco de Dados (TZ)
Professor: Clodoveu Augusto Davis Junior
Data: 30/11/2023

DICIONÁRIO DE DADOS

Estado

Coluna	Descrição	Tipo	Restrição
sigla	Sigla do estado	Varchar(5)	PK
nome	Nome do estado	Varchar(50)	NOT NULL
regiao	Região do estado	Varchar(20)	NOT NULL

PopulacaoDosEstados

Coluna	Descrição	Tipo	Restrição
sigla_uf	Sigla do estado	Varchar(5)	FK (Estado)
ano	ano	Int	PK
populacao	quantidade de habitantes	Int	NOT NULL

FocosDeCalor

Coluna	Descrição	Tipo	Restrição
sigla_estado	Sigla do estado	Varchar(5)	FK (Estado)
ano	Ano de ocorrência dos focos de calor	Int	FK (PopulacaoDosEstados)
mes	Mês de ocorrência dos focos de calor	Varchar(20)	PK
numero_focos	Quantidade de focos por mês	Int	NOT NULL
periodo	Período inicial em que os dados foram registrados	Varchar(20)	NOT NULL

Alunos: Carla Beatriz Ferreira, Gabriele Pinheiro Sá, João Marcos Ribeiro Tolentino, Manuela Monteiro Fernandes de Oliveira e Vitor Terra Mattos do Patrocínio Veloso
Disciplina: Introdução a Banco de Dados (TZ)
Professor: Clodoveu Augusto Davis Junior
Data: 30/11/2023

2. DEFINIÇÃO DOS OBJETIVOS

O objetivo do trabalho foi verificar se há alguma relação entre focos de calor e a quantidade de habitantes por estado. Nesse sentido, o grupo se propôs a responder a questão: há mais focos de calor em locais mais populosos? Para isso, o grupo averiguou a hipótese de as duas variáveis estarem relacionadas em algum grau, sendo um crescimento proporcional entre elas uma evidência disso.

3. ANÁLISE DESCRITIVA

Fazer a soma dos focos de todos os meses ja q a população é mostrada por ano e não por mês

Durante a manipulação dos bancos de dados, foi percebido que o *dataset* dos focos de calor foi aquele que abrangeu o menor período de tempo: enquanto o banco de dados populacionais têm registros de 1991 até 2021, aquele vai de 1998 até 2017. Além disso, o *dataset* a respeito dos focos de calor foi aquele que mais precisou de ser alterado, devido à má formatação e organização de seus dados.

Abaixo, os gráficos representativos de algumas consultas feitas.

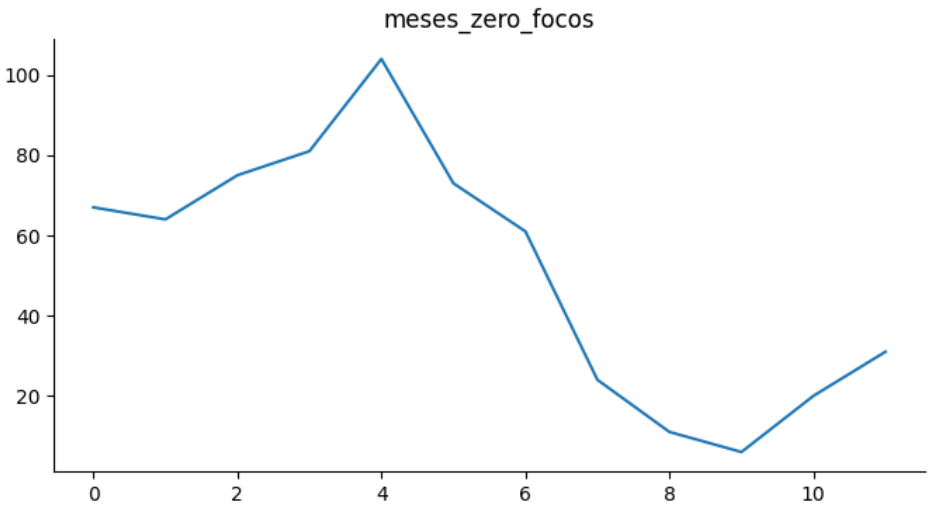

Gráfico 1: Meses sem focos de calor no período analisado (1998-2017), sendo eixo X os meses do ano e o eixo Y, a quantidade de locais que não tiveram focos de calor.

Alunos: Carla Beatriz Ferreira, Gabriele Pinheiro Sá, João Marcos Ribeiro Tolentino, Manuela Monteiro Fernandes de Oliveira e Vitor Terra Mattos do Patrocínio Veloso
Disciplina: Introdução a Banco de Dados (TZ)
Professor: Clodoveu Augusto Davis Junior
Data: 30/11/2023

	Estado	ano	media_focos
0	Sergipe	2000	1.416667
1	Sergipe	1998	1.666667
2	Roraima	1998	1.750000
3	Sergipe	2001	2.000000
4	Distrito Federal	1999	3.833333
...	...	...	...
535	Mato Grosso	2007	4201.500000
536	Mato Grosso	2005	4304.166667
537	Mato Grosso	2002	4630.166667
538	Para	2017	5433.727273
539	Mato Grosso	2004	6417.833333

Tabela 1 - Valor médio de focos por ano em cada estado ordenado por número de focos (de 1,4 a 6417,8)

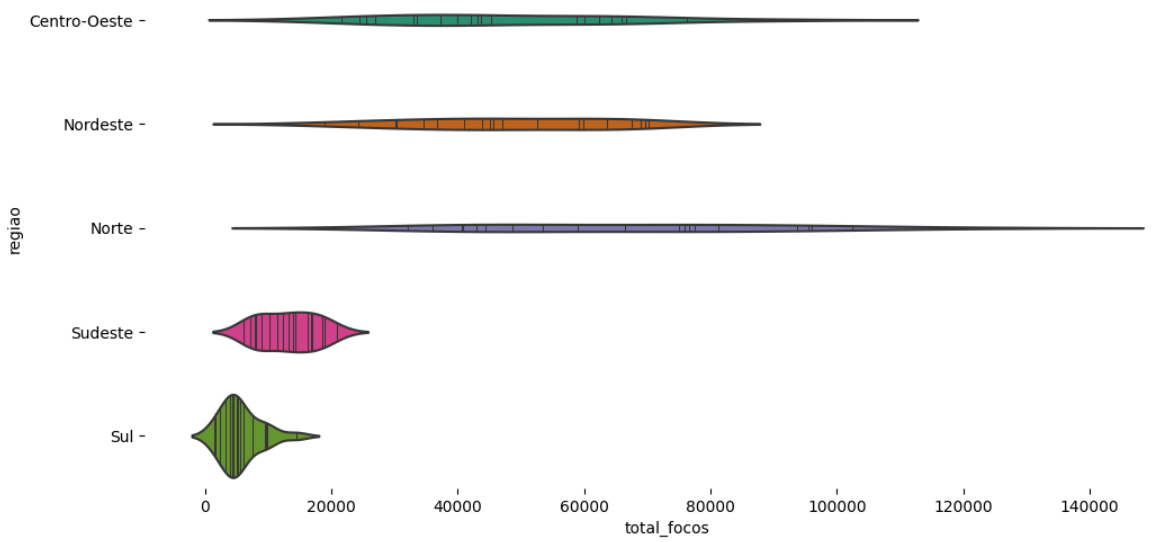

Figura 1 - Distribuição de total de focos (soma) por ano em cada região do Brasil

Alunos: Carla Beatriz Ferreira, Gabriele Pinheiro Sá, João Marcos Ribeiro Tolentino, Manuela Monteiro Fernandes de Oliveira e Vitor Terra Mattos do Patrocínio Veloso
Disciplina: Introdução a Banco de Dados (TZ)
Professor: Clodoveu Augusto Davis Junior
Data: 30/11/2023

Gráfico 2: Média do número de focos no decorrer dos anos

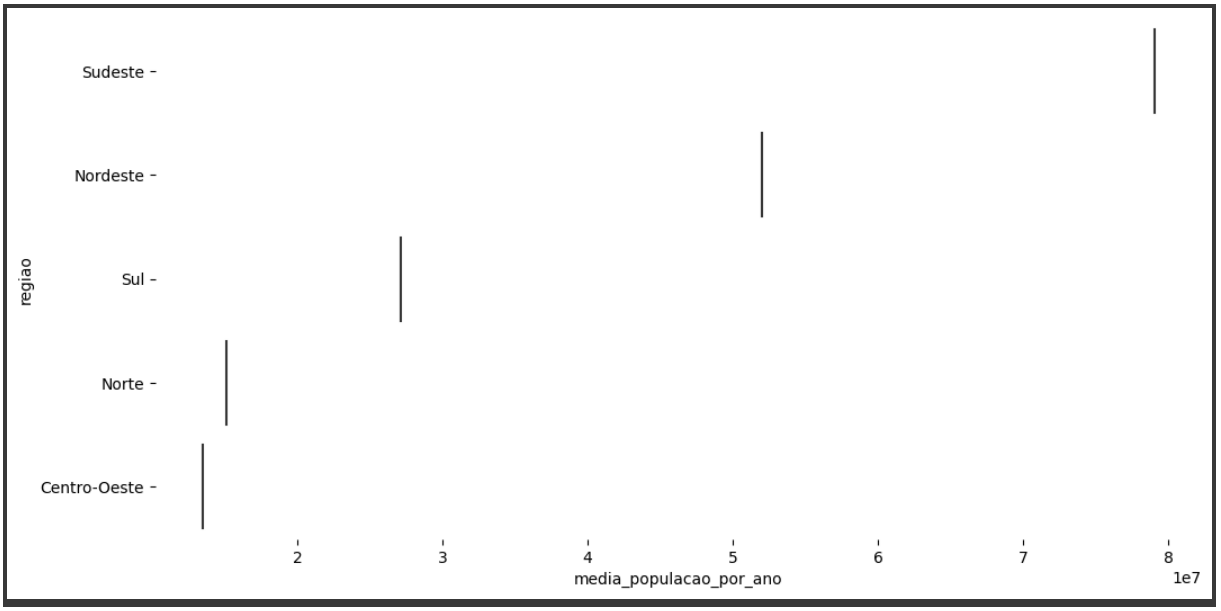

Gráfico 3: Média da população por ano por região

Alunos: Carla Beatriz Ferreira, Gabriele Pinheiro Sá, João Marcos Ribeiro Tolentino, Manuela Monteiro Fernandes de Oliveira e Vitor Terra Mattos do Patrocínio Veloso
Disciplina: Introdução a Banco de Dados (TZ)
Professor: Clodoveu Augusto Davis Junior
Data: 30/11/2023

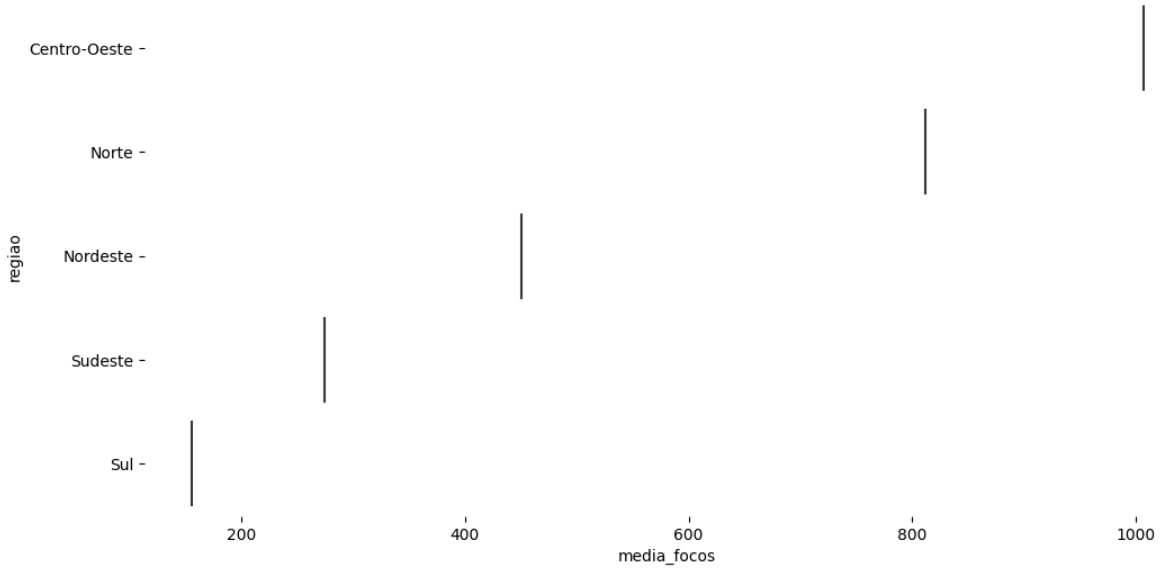

Gráfico 4: Média de focos de calor por ano por região

4. IDENTIFICAÇÃO DE VALORES DISCREPANTES

A partir das análises dos gráficos gerados e das tabelas selecionadas foi possível perceber diversos valores discrepantes, tanto para população como para focos de calor. Uma maneira útil de se observar valores discrepantes é calcular o desvio padrão dos dados coletados, para isso não foi possível utilizar os comandos em SQL dado a limitação do Notebook compartilhado (escolhido para melhor organização do grupo em um trabalho mais distribuído), portanto foi calculado utilizando múltiplas consultas alinhadas em SQL e também a partir da importação dos arquivos .csv no Google Planilhas e utilizando as fórmulas do mesmo, como apresentado nas imagens abaixo.

E	F
POR ESTADO	DESVIO PADRAO
AC	153902,8399
AL	287699,9739
AM	689342,8434
AP	184945,1398
BA	1067423,547
CE	943810,0986
DF	489718,0158
ES	482244,4497
GO	986870,7956
MA	738896,9143
MG	1839105,399
MS	331076,1199
MT	468446,1392
PA	1201128,258
PB	284211,2204
PE	840211,0777
PI	232808,4348
PR	999660,3819
RJ	1479189,952
RO	211272,5304
RR	5971508,676
RS	720578,7341
SC	879785,3373
SE	271586,8048
SP	4649794,389
TO	213246,7909

Figura 1 - Desvio padrão das populações por estados, calculado no Google Planilhas

Alunos: Carla Beatriz Ferreira, Gabriele Pinheiro Sá, João Marcos Ribeiro Tolentino, Manuela Monteiro Fernandes de Oliveira e Vitor Terra Mattos do Patrocínio Veloso
Disciplina: Introdução a Banco de Dados (TZ)
Professor: Clodoveu Augusto Davis Junior
Data: 30/11/2023

	sigla	ano	SQRT(SUM(media_menos_xi_quadrado) / COUNT(*))
0	AC	1998	139.921307
1	AL	1998	10.667969
2	AM	1998	111.855437
3	AP	1998	40.765249
4	BA	1998	869.918591
...	...	...	...
535	RS	2017	303.881094
536	SC	2017	266.555128
537	SE	2017	9.980146
538	SP	2017	813.581126
539	TO	2017	3005.650401

540 rows × 3 columns

Figura 2 - Desvio padrão dos focos de calor de cada estado

	A	B	C	D	E	F	G
1	Ano	Estado	Mes	Numero	Periodo		DESVIO PADRAO
2	1998	AC	Janeiro	0	01/01/1998		2974,15499
3	1999	AC	Janeiro	0	01/01/1999		
4	2000	AC	Janeiro	0	01/01/2000		
5	2001	AC	Janeiro	0	01/01/2001		
6	2002	AC	Janeiro	0	01/01/2002		
7	2003	AC	Janeiro	10	01/01/2003		

Figura 3 - Desvio padrão geral de focos de calor de todos os dados calculado no google planilhas

Além disso, é possível observar mais valores outliers ao se considerar a quantidade de meses que possuem valores zero como número de focos e ao observar os valores máximos registrados (de até 25963 focos em um único mês), como dispostos e analisados nas figuras e gráficos abaixo.

Alunos: Carla Beatriz Ferreira, Gabriele Pinheiro Sá, João Marcos Ribeiro Tolentino, Manuela Monteiro Fernandes de Oliveira e Vitor Terra Mattos do Patrocínio Veloso
Disciplina: Introdução a Banco de Dados (TZ)
Professor: Clodoveu Augusto Davis Junior
Data: 30/11/2023

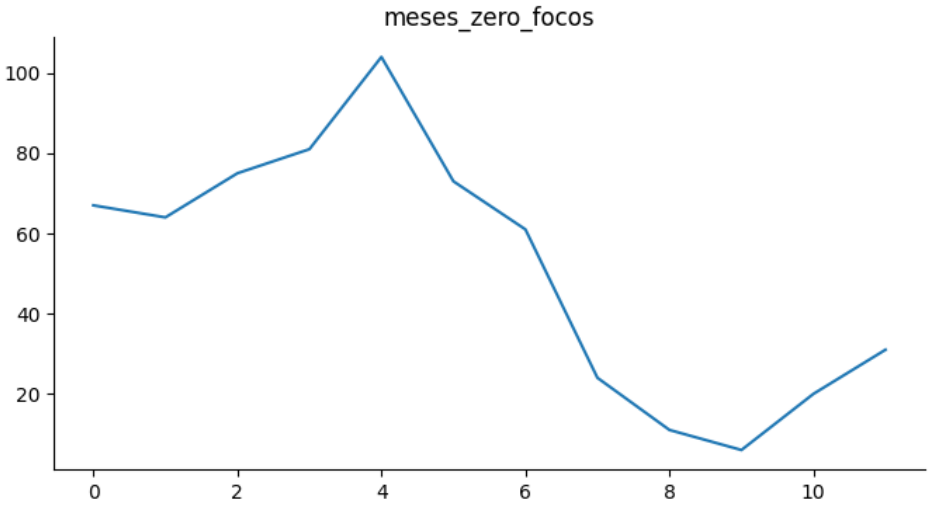

Gráfico 5: Meses sem focos de calor no período analisado (1998-2017), sendo eixo X os meses do ano e o eixo Y, a quantidade de locais que não tiveram focos de calor.

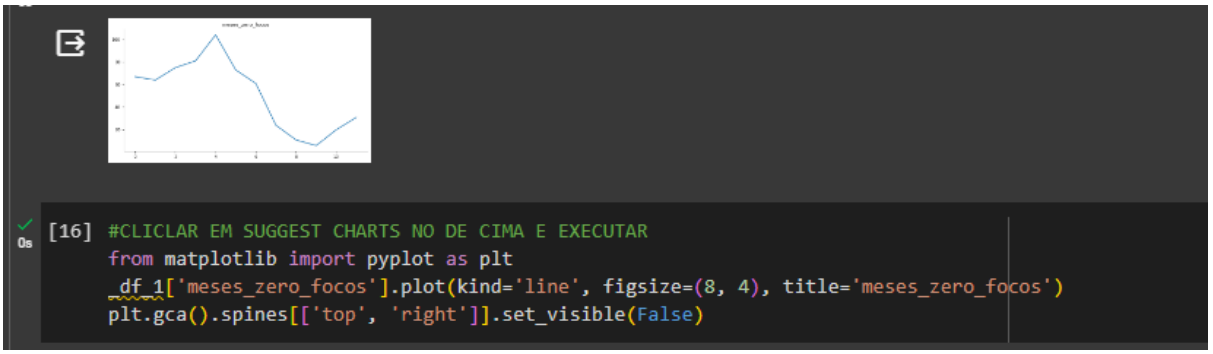

Figura 4 - Código para gerar o gráfico sugerido

	sigla_uf	mes	ano	numero_focos	regiao
0	MT	Setembro	2007	25963	Centro-Oeste
1	PA	Setembro	2017	25004	Norte
2	MT	Setembro	2004	24994	Centro-Oeste
3	MT	Setembro	2017	23945	Centro-Oeste
4	MT	Setembro	2005	20551	Centro-Oeste
5	MT	Agosto	1999	18566	Centro-Oeste
6	MT	Setembro	2010	18366	Centro-Oeste
7	PA	Agosto	2010	18130	Norte
8	MT	Setembro	2003	15790	Centro-Oeste
9	PA	Agosto	2002	15664	Norte

Figura 5 - 10 maiores focos de calor geral

5. ANÁLISE DA CORRELAÇÃO

Durante as consultas nos bancos de dados, pode-se perceber que a hipótese inicial – a de que havia uma relação de crescimento proporcional entre os focos de calor e o número de habitantes por estado – não foi confirmada. Isso porque, como é possível de observar nos gráficos 3 e 4, essa correspondência não ocorre para a região do Centro-Oeste: apesar de ser uma das regiões com menor população média, é a que possui maiores quantidades de focos de calor. Enquanto isso, a região Sudeste, que possui a maior população média, está na penúltima posição em número de focos médios.

Nesse sentido, o grupo acredita que seria necessário analisar outros critérios, como atividades econômicas e práticas sociais, para conseguir chegar a uma razão para as diferentes quantidades de focos de calor por região.

6. CONCLUSÕES

Durante a realização do Trabalho Prático, foi possível aprender a como encontrar diferentes bancos de dados, bem como extrair informações deles e relacioná-los.

Assim, o grupo pôde perceber a importância dos dados públicos estarem atualizados e disponíveis para todos, pois somente assim foi possível testar as hipóteses feitas e chegar em resultados que refletem a realidade da sociedade brasileira.

7. BIBLIOGRAFIA E BANCOS UTILIZADOS

Bancos de Dados Utilizados:

- Banco de dados de focos de calor por estados, pelo Sistema Nacional de Informações Florestais - SNIF, disponível em: https://dados.gov.br/dados/conjuntos-dados/snif. Atualizado em 03 de março de 2022. Acessado pela última vez em 30 de novembro de 2023.
- Banco de dados de população brasileira, pelo Instituto Brasileiro de Geografia e Estatística - IBGE, disponível em: https://basedosdados.org/dataset/d30222ad-7a5c-4778-a1ec-f0785371d1ca?table=2440d076-8934-471f-8cbe-51faae387c66. Atualizado em 2022 (estimativa). Acessado pela última vez em 30 de novembro de 2023.

Outros links importantes:

- Repositório no github: https://github.com/carlabferreira/TP2---IBD
- Planilha compartilhada no Google Planilhas para cálculo do desvio padrão de focos de calor: focosdecalornormalizado
- Planilha compartilhada no Google Planilhas para cálculo do desvio padrão das populações: populacao1991-2021
- Notebook compartilhado com Python e consultas SQL utilizadas para análise: TP2.ipynb