Trabalho: Relatório Final - Definição de Dataset, EDA e Modelagem
Aluno: Emerson Adam
Disciplina: CID22-IntroProgrCiênciaDeDados-202301
Professor: Luiz Celso Gomes Jr.

Análise da Base de dados abertos da SiGesGuarda de Curitiba

1. Introdução

Neste trabalho proposto na disciplina de Introdução à Programação Para Ciência De Dados, do Curso de Especialização em Ciência de Dados e suas Aplicações promovido pela UTFPR, tem como objetivo a definição de um dataset de dados abertos da prefeitura de Curitiba.

O dataset escolhido foi o SiGesGuarda, que contempla todos os registros entre 2009 e 2023, contendo 429347 linhas e 35 colunas com diversas informações. Link para o dataset

Este conjunto de dados será modificado e limpo, através da linguagem de programação Python e bibliotecas Pandas e matplotlib. Após os ajustes será analisado de forma exploratória, encontrando informações de valor para tomada de decisão, como opção, das autoridades competentes.

2. Sobre o problema

Após a definição do Dataset, foi proposto pelo professor Luiz realizarmos alguns questionamentos que podem ser respondidos através da análise dos dados. Com isso podemos confirmar de fato alguns pontos ao invés da suposição.

Portanto as perguntas que serão respondidas na fase de exploração serão:

- Quais são os tipos de registros mais comuns em Curitiba?
- Quais os dias da semana possuem mais registros?
- Quais bairros possuem mais registros?
- Qual a distribuição das ocorrências nos anos?
- Qual o período com maior frequência de registros referente à perturbação de sossego?
- Quais os dias em que há maior registros referentes à perturbação de sossego, evidenciando os períodos?

Já na **modelagem**, veremos mais abaixo que há um aumento significativo de registros entre os anos de 2019 até 2023. Portanto:

- Existe relação entre o período da pandemia Covid-19 e o aumento de registros?

3. Análise do dataset - EDA

Os passos desta análise podem ser verificados no documento enviado anexo em HTML com o título *Trabalho Definição de Dataset Limpeza e Análise dos Dados - Emerson Adam - CDA – UTFPR.*

Esta etapa consiste em importar os dados, realizar a apropriação do contexto, entendendo melhor seu conjunto de dados, realizar a limpeza e a criação de gráficos para encontrar as respostas para as perguntas acima, definidas no tópico 2.

Trabalho: Relatório Final - Definição de Dataset, EDA e Modelagem
Aluno: Emerson Adam
Disciplina: CID22-IntroProgrCiênciaDeDados-202301
Professor: Luiz Celso Gomes Jr.

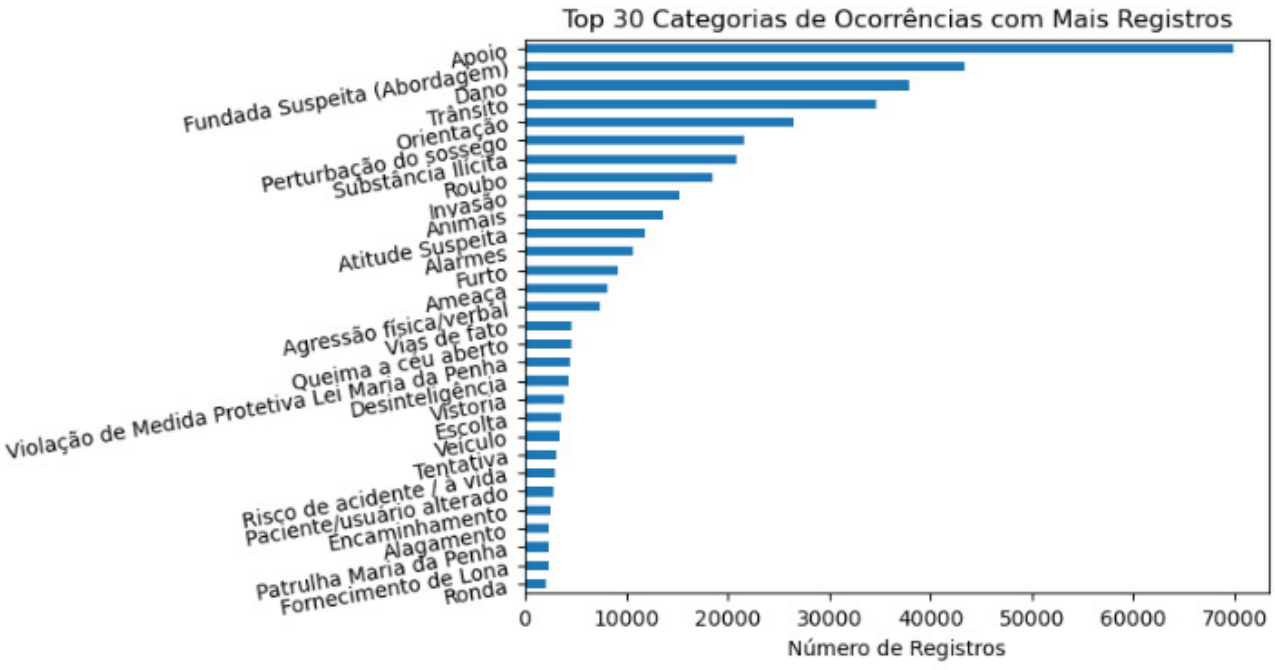
Este dataset contém um dicionário de dados, porém, como o dataset foi alterado e selecionado somente algumas colunas, será colocado neste documento somente as que foram utilizadas na análise. Link do documento original.

DICIONÁRIO DE DADOS DO DATASET SIGESGUARDA			
COLUNAS	**TIPO**	**TAMANHO**	**DESCRIÇÃO DOS CAMPOS**
ATENDIMENTO_BAIRRO_NOME	varchar	20	Nome do bairro em que foi realizado o atendimento
LOGRADOURO_NOME	varchar	70	Nome do logradouro
NATUREZA1_DESCRICAO	varchar	100	Descrição da primeira natureza cadastrada na ocorrência
OCORRENCIA_ANO	int	-	Ano de cadastro da ocorrência
OCORRENCIA_CODIGO	int	-	Código da ocorrência
OCORRENCIA_DATA	datetime	-	Data da ocorrência
OCORRENCIA_DIA_SEMANA	varchar	20	Dia da semana em que a ocorrência foi cadastrada
OCORRENCIA_HORA	varchar	8	Hora em que a ocorrência foi cadastrada
OCORRENCIA_MES	int	-	Mês em que a ocorrência foi cadastrada
OPERACAO_DESCRICAO	varchar	70	Nome da operação que realizará o atendimento, caso haja
ORIGEM_CHAMADO_DESCRICAO	varchar	70	Local em que se originou a chamada

Após a apropriação, podemos começar a responder as perguntas.

- **Quais são os tipos de registros mais comuns em Curitiba?**

Para responder esta pergunta, foi efetuado um gráfico com os 30 assuntos com mais registros, conforme abaixo:



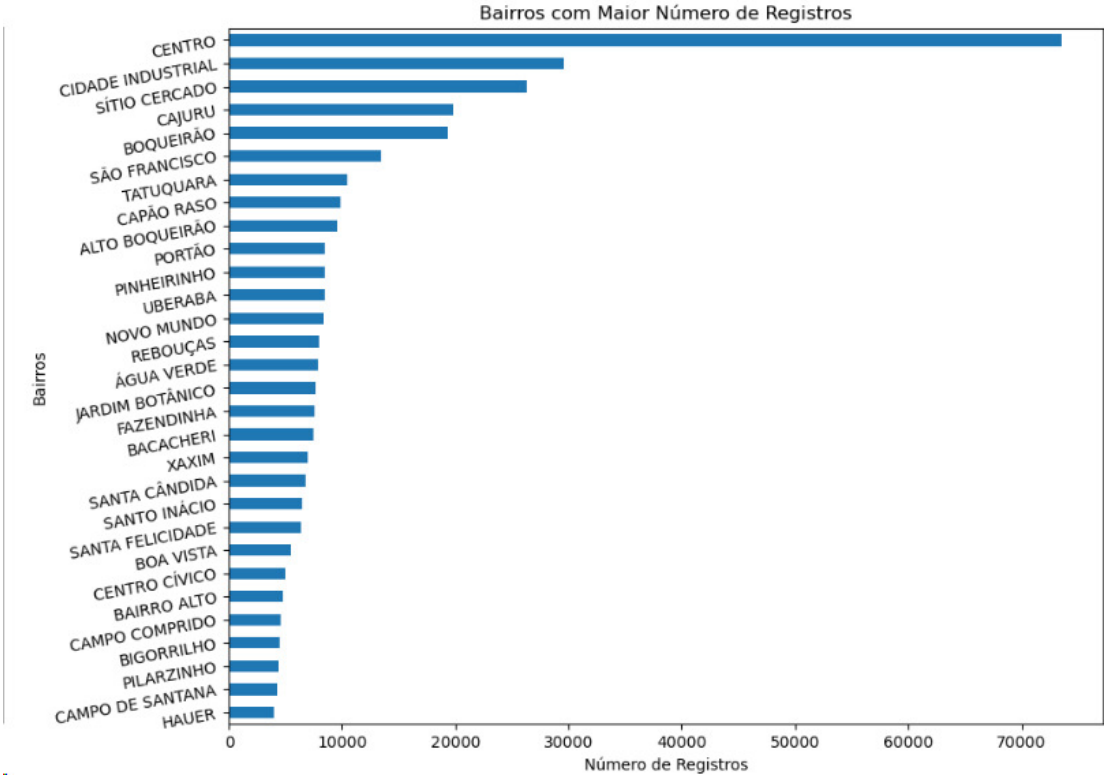
- **Quais os dias da semana possuem mais registros?**

  Para esta pergunta, realizamos o gráfico abaixo:

Identifica-se que os 3 dias com mais chamados são no fim de semana a partir de sexta-feira.

- **Quais bairros possuem mais registros?**

Para responder esta pergunta, o gráfico abaixo foi gerado, contendo os 30 maiores bairros em relação ao número de registros.

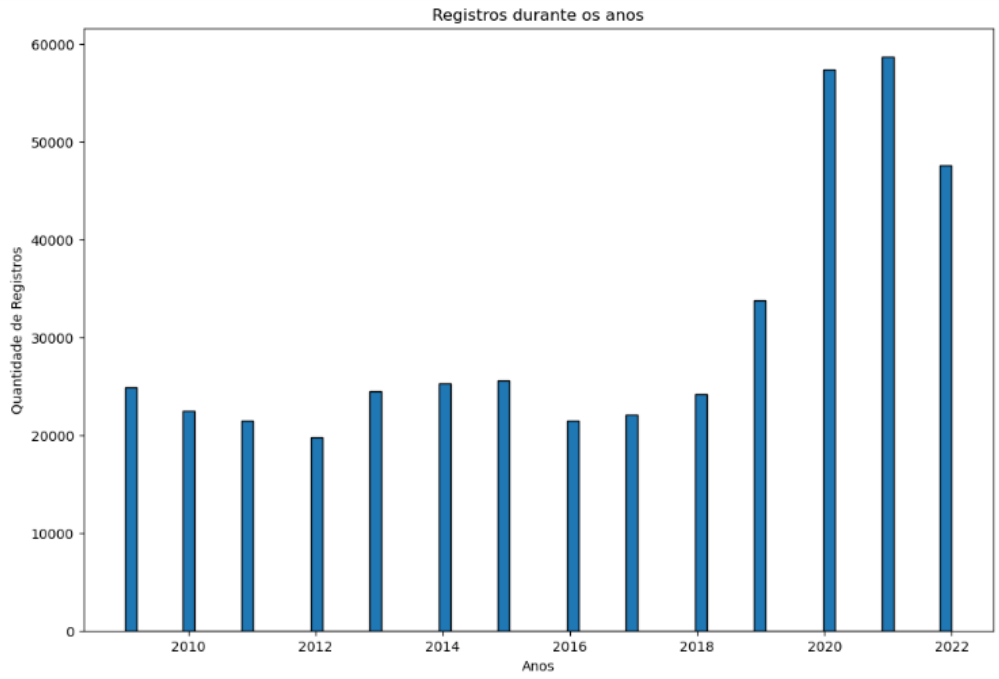


Nota-se que Centro, CIC e Sitio Cercado possuem uma grande diferença entre os demais.

- **Qual a distribuição das ocorrências nos anos?**

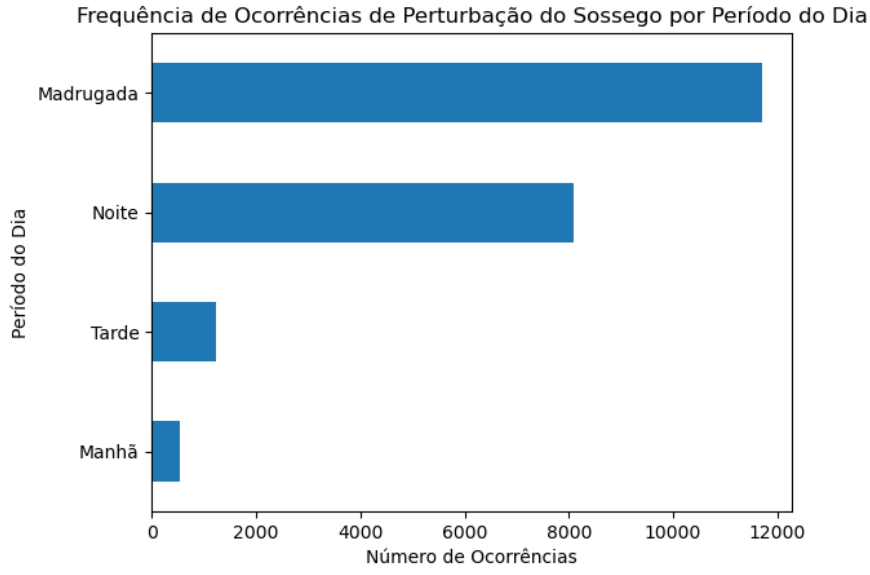
Para responder esta pergunta, foi gerado o gráfico no estilo histograma abaixo:

Nota-se um aumento considerável de registros a partir de 2019, gráfico que levou a **hipótese ser considerada.**



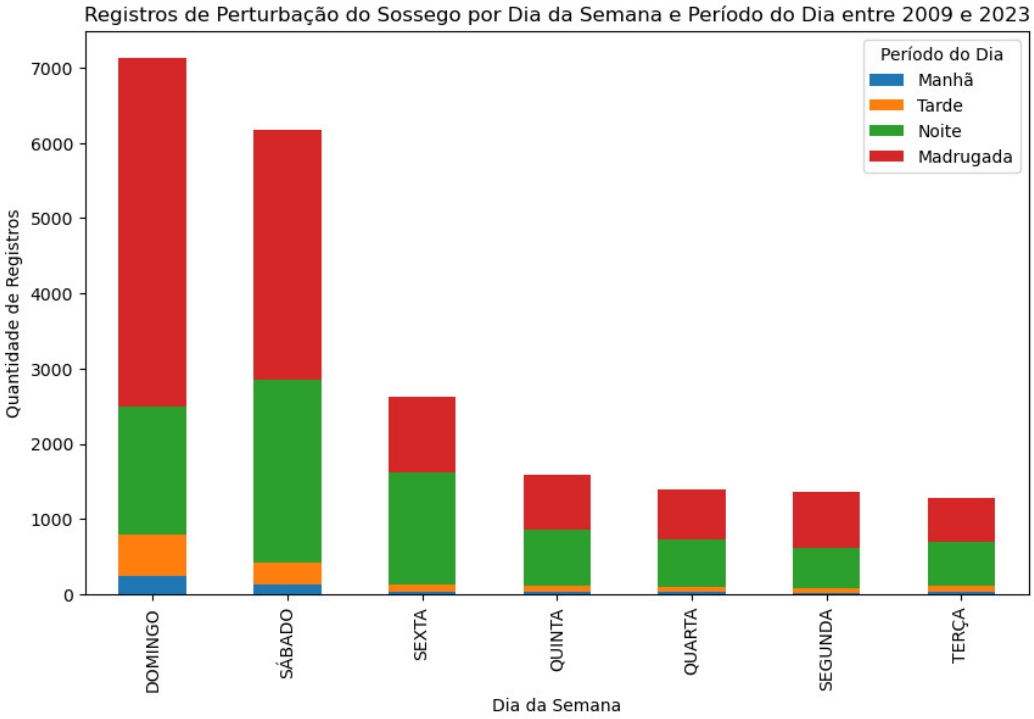
- **Qual o período com maior frequência de registros referente à perturbação de sossego?**

Para responder esta questão, geramos o gráfico de barras abaixo, separando por período de acordo com a hora da ocorrência, nota-se que de madrugada e a noite são os períodos que possuem maiores registros



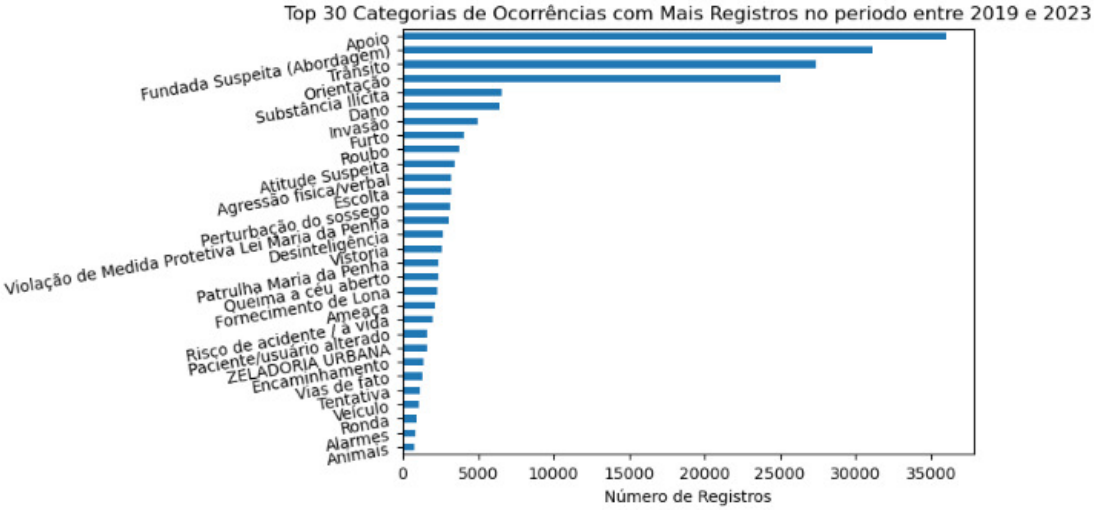
- **Quais os dias em que há maior registros referentes à perturbação de sossego, evidenciando os períodos?**

Para responder esta pergunta, geramos o gráfico abaixo onde fica evidente que ao final de semana, a partir da noite de sexta-feira até o início da manhã de segunda-feira temos um aumento significativo dos registros de perturbação do sossego. Análise efetuada entre os anos de 2009 e 2023.

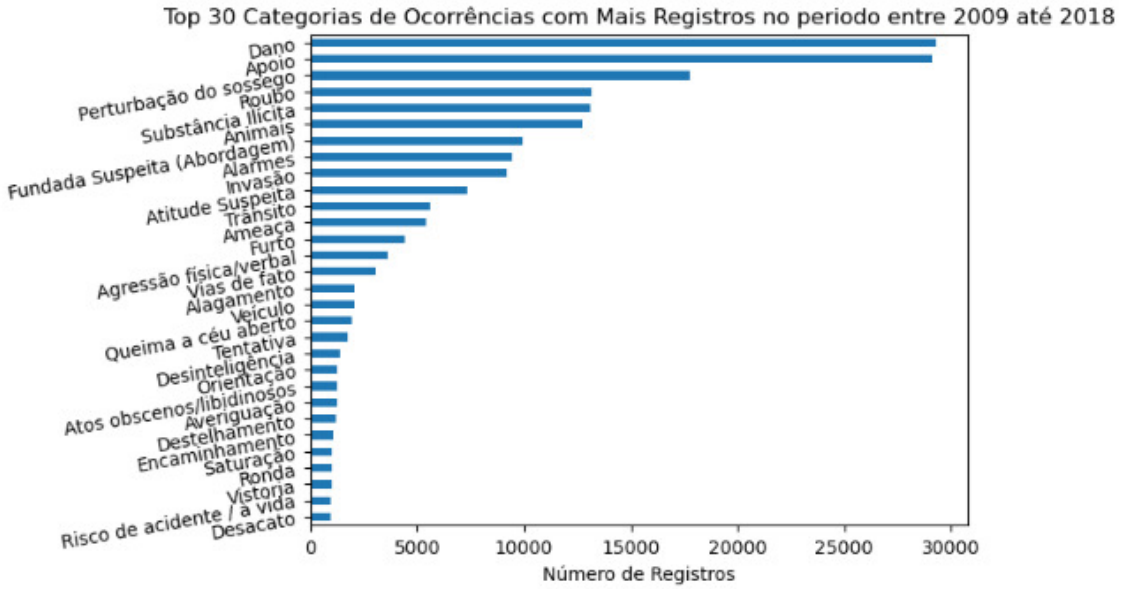


4. Modelagem

Para responder a hipótese levantada sobre a existência de relação entre o período da pandemia Covid-19 e o aumento de registros, teremos que separar em 2 grupos, um pré covid e outro pós. Com isso podemos observar que a categoria apoio aparece em primeiro lugar no dataframe pós covid, que contempla registros a partir de 2019.



Já no período pré covid, a categoria apoio aparece em segundo lugar



Para realizar um teste estatístico especifico na categoria apoio para tentar responder a hipótese de que a pandemia covid-19 teve participação no aumento de registros, precisamos filtrar as subcategorias, do grupo apoio, que são relacionadas à saúde. Portanto temos as subcategorias abaixo:

```
[23]: #verificando quais as subcategorias da natureza apoio, pois é a categoria que mais tem aumento de todos os registros
df_apoio = df[df['NATUREZA1_DESCRICAO'] == 'Apoio']
#print(df_apoio)
df_apoio['SUBCATEGORIA1_DESCRICAO'].unique()
```

```
t[23]: array(['Apoio ao cidadão - PRESTAÇÃO DE SOCORRO/SALVAMENTO',
       'Apoio ao cidadão - ORIENTAÇÃO', 'Apoio à outros órgãos',
       'Apoio ao GM', 'Apoio à SMS', 'Apoio ao SIATE', 'Apoio à PMPR',
       'Apoio à URBS', 'Apoio à FAS', 'Apoio ao SAMU',
       'Apoio ao Conselho Tutelar', 'Apoio à SMAD', 'Apoio à SMU',
       'Apoio ao Patrimônio', 'Apoio à Polícia Civil', 'Apoio à SME',
       'Apoio à SMEL', 'Apoio à SETRAN', 'Apoio à COHAB', 'Apoio à SMMA',
       'Apoio à Vigilância Sanitária', 'Apoio à SMOP', 'Apoio à COSEDI',
       'Apoio à Polícia Federal', 'Apoio à SMAB', 'Apoio à SAM',
       'Apoio ao IML', 'Apoio ao Corpo de Bombeiros',
       'Apoio à Polícia Rodoviária Federal', 'Apoio ao Horto Municipal',
       nan, 'Atendimento a pessoa em situação de vulnerabilidade social',
       'Apoio ao cidadão - PCD Pessoa Com Deficiencia'], dtype=object)
```

E analisando os nomes, iremos separar somente as categorias abaixo, separando em 2 dataframes, pré e pós covid:

Trabalho: Relatório Final - Definição de Dataset, EDA e Modelagem
Aluno: Emerson Adam
Disciplina: CID22-IntroProgrCiênciaDeDados-202301
Professor: Luiz Celso Gomes Jr.

```
In [25]: #agora vamos olhar somente os registros de apoio relacionados à saúde
naturezas_interesse = ['Apoio ao cidadão - PRESTAÇÃO DE SOCORRO/SALVAMENTO', 'Apoio à SMS', 'Apoio ao SIATE', 'Apoio ao SAMU']
df_filtro_posCovid = df_filtro_posCovid[df_filtro_posCovid['SUBCATEGORIA1_DESCRICAO'].isin(naturezas_interesse)]
df_filtro_posCovid
```

Out[25]:

	ATENDIMENTO_BAIRRO_NOME	LOGRADOURO_NOME	NATUREZA1_DESCRICAO	SUBCATEGORIA1_DESCRICAO	OCORRENCIA_ANO	OCORRENCIA_
231793	NOVO MUNDO	DOM BOSCO	Apoio	Apoio ao SAMU	2019	
231794	SANTO INÁCIO	BR-277 - CURITIBA / PONTA GROSSA	Apoio	Apoio ao cidadão - PRESTAÇÃO DE SOCORRO/SALVAM...	2019	
231795	CIDADE INDUSTRIAL	SEN. ACCIOLY FILHO	Apoio	Apoio à SMS	2019	
231796	CIDADE INDUSTRIAL	SEN. ACCIOLY FILHO	Apoio	Apoio à SMS	2019	
231800	CIDADE INDUSTRIAL	SEN. ACCIOLY FILHO	Apoio	Apoio à SMS	2019	
...	...	...	...	...	...	
429265	BATEL	BISPO DOM JOSÉ	Apoio	Apoio ao SAMU	2022	

```
In [26]: #agora os registros preCovid
df_filtro_preCovid = df_filtro_preCovid[df_filtro_preCovid['SUBCATEGORIA1_DESCRICAO'].isin(naturezas_interesse)]
df_filtro_preCovid
```

Out[26]:

	ATENDIMENTO_BAIRRO_NOME	LOGRADOURO_NOME	NATUREZA1_DESCRICAO	SUBCATEGORIA1_DESCRICAO	OCORRENCIA_ANO	OCORRENCIA_
11	TABOÃO	OSWALDO MACIEL	Apoio	Apoio ao cidadão - PRESTAÇÃO DE SOCORRO/SALVAM...	2009	

Com esses 2 dataframes conseguimos contar separadamente a quantidade de registros relacionados ao apoio referentes à saúde nos dois períodos, pré e pós pandemia. Gerando assim variáveis discretas para a realização de teste T e teste de Mann-Whitney U.

Fazendo uma média simples vemos que em 9 anos referente ao período pré pandemia, temos 13378 linhas, o que da uma média de 1486 chamados por ano, já no período pós pandemia temos 11904 registros em 4 anos, o que dá uma média de 2976 chamados por ano, praticamente 50% a mais no período da pandemia, isso somente para registros relacionados ao apoio à serviços de saúde.

Ao realizar o teste T e o teste de Mann-Whitney, encontramos um valor de P muito alto o que nos diz que não podemos confiar estatisticamente para que isso seja uma hipótese válida.

```
In [30]: # contando numero de eventos no DataFrame df_filtro_preCovid
contagem_preCovid = df_filtro_preCovid['SUBCATEGORIA1_DESCRICAO'].value_counts()

# para poscovid
contagem_posCovid = df_filtro_posCovid['SUBCATEGORIA1_DESCRICAO'].value_counts()

pre = contagem_preCovid.values
pos = contagem_posCovid.values

# teste T
t_statistic, p_value = stats.ttest_ind(pre, pos)

print("teste T:", t_statistic)
print("Valor de p:", p_value)
```

```
teste T: 0.16531601503127677
Valor de p: 0.8741246352632057
```

```
In [31]: # teste de Mann-Whitney U
statistic, p_value = stats.mannwhitneyu(pre, pos)
print("teste de Mann-Whitney U:", statistic)
print("Valor de p:", p_value)
```

```
teste de Mann-Whitney U: 9.0
Valor de p: 0.8857142857142857
```

5. Conclusão

Após análises, podemos considerar que, especificamente na categoria perturbação de sossego, da base da SiGesGuarda, há um volume maior de registros entre a noite de sexta-feira até o início da manhã de segunda.

Portanto ações devem ser recomendadas para que esse número diminua, como por exemplo uma maior fiscalização e punição para quem esteja infringindo a lei e medidas educativas que relevem o respeito para com todos.

Com relação à hipótese de que a pandemia de covid-19 tenha aumentado os registros de apoio da Guarda Municipal de Curitiba entre os anos 2019~2023, não podemos afirmar estatisticamente de que esta hipótese é válida, pois apesar de haver um aumento de 50% de registros deste tipo num período de 4 anos contra 9 anos pré pandemia, o valor elevado de p dos testes estatísticos mostram que não podemos confiar nos dados. Isto também reflete no tipo de cadastro efetuado, não tendo mais detalhes com relação à apoios específicos para a pandemia.

Para trabalhos futuros, seria necessário mais detalhes nos dados, categorizando melhor os atendimentos possibilitando uma melhor análise relacionada aos atendimentos prestados durante a pandemia.