



**Gabriel da Silva Gonçalves**

Construção de um pipeline de dados envolvendo busca, coleta,  
modelagem, carga e análise dos dados.  
Análise sobre tiroteios policiais fatais nos Estados Unidos desde 2015, da  
base de dados “Fatal Encounters”

PROJETO DE PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E ANALYTICS  
APRESENTADO AO DEPARTAMENTO RESPONSÁVEL  
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO  
DO DIPLOMA DE PÓS-GRADUAÇÃO LATO SENSU

Rio de Janeiro, 01 de Outubro de 2023.

## 1. Busca pelos dados

A busca pela base de dados foi feita no [Kaggle](#).

O link é: [Fatal US Police Violence \(kaggle.com\)](#)

Na verdade, por se tratar de uma base bastante simples e que já tratada, foi optado por extrair de onde esses dados vieram, ou seja, do site: [Fatal Encounters – A step toward creating an impartial, comprehensive and searchable national database of people killed during interactions with police.](#)

O download foi feito pelo botão “Download FE Database”



Descrição da base:

“Fatal Encounters é um banco de dados de incidentes em que um indivíduo morre durante um encontro com policiais. A maioria dos encontros ocorre como resultado de homicídio policial, como quando os policiais atiram em uma pessoa que representa uma ameaça letal para eles ou outras pessoas. No entanto, existem outros tipos de encontros que não envolvem atos de homicídio policial, mas nos quais a polícia está envolvida ou presente. Os exemplos incluem um acidente de veículo durante uma perseguição policial ou um suicídio em uma situação de barricada.”

(informação retirada da planilha “FATAL ENCOUNTERS DOT ORG SPREADSHEET (See Read me tab).xlsx”)

A própria planilha oferece um dicionário de dados na aba “Read Me”, coluna “Codebook in development”.

Read me notes and caveats	"Codebook" in development
<p>This Google sheet is managed by D. Brian Burghart of Fatal Encounters Dot Org. <a href="https://fatalencounters.org">fatalencounters.org</a></p> <p>To download, go to the Google spreadsheet, link below, go under File&gt;Download as&gt; and pick your format. We recommend comma-separated values.</p> <p><a href="https://docs.google.com/spreadsheets/d/1dKmaV_JMwG8XBzRyP8b4d9C0pka7FL7mvsyzvAoE/edit#gid=0">https://docs.google.com/spreadsheets/d/1dKmaV_JMwG8XBzRyP8b4d9C0pka7FL7mvsyzvAoE/edit#gid=0</a></p> <p>Fatal Encounters documents non-police deaths that occur when police are present or are precipitated by police action or presence. Officer deaths are included when caused by another officer, including friendly fire incidents, and criminal actions—like domestic violence—and suicides that occur when other officers are present. Officer vehicle-related deaths are included when they are caused by another officer. Homicides of officers by felons or deaths in the regular course of duties are not generally documented in the database.</p> <p>We believe we include all the available records for all 50 states and DC back to 2000, but there are several data points that we think are too poorly reported in the news media to result in accurate results for analysis: disposition and mental state. Our racial data (Column D) is the best that exists, but it's pretty spotty and gets worse prior to 2013. Beginning in 2020 we added two columns that regard imputed race. We generally do weekly updates on Tuesdays (although for practical reasons, sometimes that's extended later into the week), so be aware we're usually a few days behind. Government data also suggests that police chase deaths are often not reported in news media, so our data almost certainly understates those totals.</p> <p>This data is available for anyone to use for whatever purpose they choose. The only requirement for use is if users spot any errors, please report it to <a href="mailto:d.burghart@fatalencounters.org">d.burghart@fatalencounters.org</a>. As of October 2020, we require attribution, as some users are using the data in unethical ways, for example, we don't track mental illness, we track whether the officer knew the person was in mental crisis before they arrived. Missing this data would result in wildly inaccurate underestimates of the impact of gun violence and drug or</p>	<p>Column A: Fatal Encounters' Unique ID: Generally speaking, the UID works like other UIDs work; new IDs are added to new incidents without consideration of the date of the incident or the date of its inclusion into the dataset. However, twice since 2012, we've had to rebuild UIDs because of problems with the Google Spreadsheet. This in no way infers a problem with the sheet, more likely operator error, but in one instance, formulas crept in, and in the other, UIDs, which are generated manually, developed duplicate values. Also, if a duplicate record is discovered in the data, the duplicate is replaced with a non-duplicate, in order to keep the UIDs sequential, and to enable visual verification that the UIDs are working as intended. Column A is manually replicated in Column AA as a backup to enable replacement of Column A, if problems with the UIDs ever arise again. Fatal Encounters recommends researchers note download dates. The current sheet always includes the most up-to-date data and accurate data.</p> <p>Column B: Subject's name: Names contain all information Fatal Encounters has been able to collect, including nicknames if available. Often additional information for names comes through obituaries or social media. In case of "Names withheld by police," the names are sometimes also voluntarily withheld by news media—especially in the cases of suicide—but police fail to publish the names through public disclosures. In the cases of "aka," sometimes this indicates errors or variations of names reported in news media, and sometimes it indicates aliases used by the decedent. In instances of transgender individuals, it's a regrettable necessity because news media and police often refuse to identify transgender people by their chosen names and gender.</p> <p>Column C: Subject's age: Ages are generally reported in news media, official documents and obituaries. Ages frequently change with updates in articles, with early reporting being the least accurate, neither police nor news media using "time between date" calculators, instead subtracting birth year from current year. In cases where police and media reports and obituaries conflict, and there is no birth date available, Fatal Encounters generally goes with the age stated in the obituary.</p> <p>Column D: Subject's gender: Male, Female, Transgender, or empty cell</p> <p>Column E: Subject's race: In Column E, race is usually reported based on visual evidence or official reports. Visual evidence includes images in news stories, obituaries, or body camera, or other surveillance videos. Sometimes race is disclosed in a news article as an identifying</p>

As suas abas mais relevantes são: “Form Responses”, onde os dados de fato estão e “State Abbreviations and Populat”, em que há dados sobre a população de cada estado.

## 2. Tratamento prévio

A fim de criar arquivos csv separados, cada um com o nome de uma aba, foi criado um script para fazer isso. Esse script, no Github, é chamado de “separar\_abas.py”.

Foi feito um tratamento prévio nos dados a fim de tornar o seu upload mais fácil para a nuvem. Esse processo está melhor descrito no arquivo “analise\_inicial.ipynb”

## 3. Uso da nuvem (Azure)

Por já possuir familiaridade com a plataforma, foi optado por se utilizar o Microsoft Azure com o intuito de disponibilizar a base de dados em um banco SQL e usar o Databricks posteriormente.

Após criar uma conta free trial no Azure, foi, primeiramente, necessário criar um storage account e um resource group:

Home > Storage accounts >

## Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#)

### Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription \* Azure subscription 1

Resource group \* (New) resourcegroup1  
Create new

### Instance details

Storage account name \* gahb123456789012

Region \* (US) East US  
Deploy to an edge zone

Performance \* ☒ Standard: Recommended for most scenarios (general-purpose v2 account)  
☐ Premium: Recommended for scenarios that require low latency.

Redundancy \* Geo-redundant storage (GRS)

Review < Previous Next : Advanced >

Em seguida, dentro desse storage account, foi feito o upload dos arquivos em csv, sendo também necessário criar um container:

Overview >

Upload Open in Explorer Delete Move Refresh Open in mobile CLI / PS Feedback

Essentials

Resource group (readonly) resourcegroup1  
Location East US  
Primary/Secondary Location Primary: East US, Secondary: West US  
Subscription (readonly) Azure subscription 1  
Subscription ID gahb123456789012  
Disk state Primary: Available, Secondary: Available

Tags (edit) Add tags

Properties Monitoring Capabilities (7) Recommendations (0) Tutorials Tools + SDKs

**Blob service**

Hierarchical namespace Disabled  
Default access tier Hot  
Blob anonymous access Disabled  
Blob soft delete Enabled (7 days)  
Container soft delete Enabled (7 days)  
Versioning Disabled  
Change feed Disabled  
NFS v3 Disabled  
Allow cross-tenant replication Disabled

**File service**

Large file share Disabled  
Active Directory Not configured  
Default share-level permissions Disabled  
Soft delete Enabled (7 days)

**Security**

Require secure transfer for REST API operations  
Storage account key access  
Minimum TLS version  
Infrastructure encryption

**Networking**

Allow access from  
Number of private endpoint connections  
Network routing  
Access for trusted Microsoft services  
Endpoint type

### Upload blob

\*\*\* Uploading on blobs...  
Attempting to upload 2 blob(s)

Drag and drop files here  
or  
Browse for files

Select an existing container  
gabrielcontainer  
Create new

☒ Overwrite if files already exist

Advanced

Upload Give feedback

Current uploads

fatal\_encounters\_main.csv 0 / 17.59 MiB  
fatal\_encounters\_states\_... 2.75 KiB / 2.75 KiB

Dessa forma, os arquivos ficaram disponíveis:

gabrielcontainer

Search

Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots Create snapshot Give feedback

Authentication method: Access key (Switch to Azure AD User Account)  
Location: gabrielcontainer

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
fatal_encounters_main.csv	9/30/2023, 7:44:51 PM	Hot (inferred)		Block blob	17.59 MiB	Available
fatal_encounters_states_population.csv	9/30/2023, 7:44:48 PM	Hot (inferred)		Block blob	2.75 KiB	Available

Tendo os arquivos, partiu-se para a etapa de criação do servidor do banco de dados SQL:

Home > SQL databases > Create SQL Database >

# Create SQL Database Server

Microsoft

## Server details

Enter required settings for this server, including providing a name and location. This server will be created in the same subscription and resource group as your database.

Server name \*


serveradmin

.database.windows.net

Location \*

(US) East US

## Authentication

 Azure Active Directory (Azure AD) is now Microsoft Entra ID. [Learn more](#)

Select your preferred authentication methods for accessing this server. Create a server admin login and password to access your server with SQL authentication, select only Microsoft Entra authentication [Learn more](#) using an existing Microsoft Entra user, group, or application as Microsoft Entra admin [Learn more](#), or select both SQL and Microsoft Entra authentication.

Authentication method

☐ Use Microsoft Entra-only authentication

☐ Use both SQL and Microsoft Entra authentication

☒ Use SQL authentication

Server admin login \*

serveradmin

Password \*

\*\*\*\*\*

Confirm password \*

\*\*\*\*\*

OK

Home > SQL databases >

Create SQL Database

Microsoft

BasicsNetworkingSecurityAdditional settingsTagsReview + create

Create a SQL database with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. [Learn more](#)

Project details  
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.  

Subscription \*  
Resource group \*

Database details  
Enter required settings for this database, including picking a logical server and configuring the compute and storage resources  

Database name \*  
Server \*

Want to use SQL elastic pool?  
Workload environment

Default settings: provided for Production workloads. Configurations can be modified as needed.

Review + createNext : Networking >

Cost summary

General Purpose (GP_Gen5_2)	
Cost per vCore (in USD)	184.09
vCores selected	x 2
Cost per GB (in USD)	0.12
Max storage selected (in GB)	x 416
ESTIMATED COST / MONTH	372.97 USD

Com isso, foi criado um Data Factory para, posteriormente, ser realizado o processo de pipeline:

The screenshot shows the 'Create Data Factory' wizard in the Azure portal. The breadcrumb navigation at the top reads: Home > Create a resource > Marketplace > Data Factory >. The main heading is 'Create Data Factory' with a three-dot menu icon. Below the heading are tabs for 'Basics', 'Git configuration', 'Networking', 'Advanced', 'Tags', and 'Review + create'. A message states: 'One-click to create data factory with sample pipeline and datasets. [Try it](#)'. The 'Project details' section instructs: 'Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.' It contains two dropdown menus: 'Subscription' (selected: 'Azure subscription 1') and 'Resource group' (selected: 'resourcegroup1', with a 'Create new' link below it). The 'Instance details' section contains three dropdown menus: 'Name' (selected: 'gabrielfactory'), 'Region' (selected: 'East US'), and 'Version' (selected: 'V2'). At the bottom are three buttons: 'Previous', 'Next', and 'Review + create'.

Ao acessar o Data Factory, a etapa de criação do pipeline foi realizada e envolveu: ingestão das tabelas presentes no blob storage, criação do dataflow e do pipeline.

The screenshot shows the 'New linked service' dialog in the Azure Data Factory interface. The left sidebar shows 'Factory Resources' with 'Pipelines' (1), 'Change Data Capture (preview)' (0), 'Datasets' (0), 'Data flows' (0), and 'Power Query' (0). The 'Activities' pane lists various tasks like 'Move and transform', 'Synapse', 'Azure Data Explorer', etc. The main area is titled 'New linked service' for 'Azure Blob Storage'. Fields include: 'Name' (AzureBlobStorage1), 'Description' (empty), 'Connect via integration runtime' (AutoResolveIntegrationRuntime), 'Authentication type' (Account key), and 'Connection string' (Azure Key Vault). Under 'Account selection method', 'From Azure subscription' is selected, showing 'Azure subscription 1' and 'Storage account name 1' (both redacted). At the bottom are 'Create' and 'Cancel' buttons, and a 'Test connection' link.

## Set properties

Name

DelimitedText1

Linked service \*

AzureBlobStorage1

File path

gabrielcontainer

/ Directory

/ fatal\_encounters\_main....

First row as header



Import schema

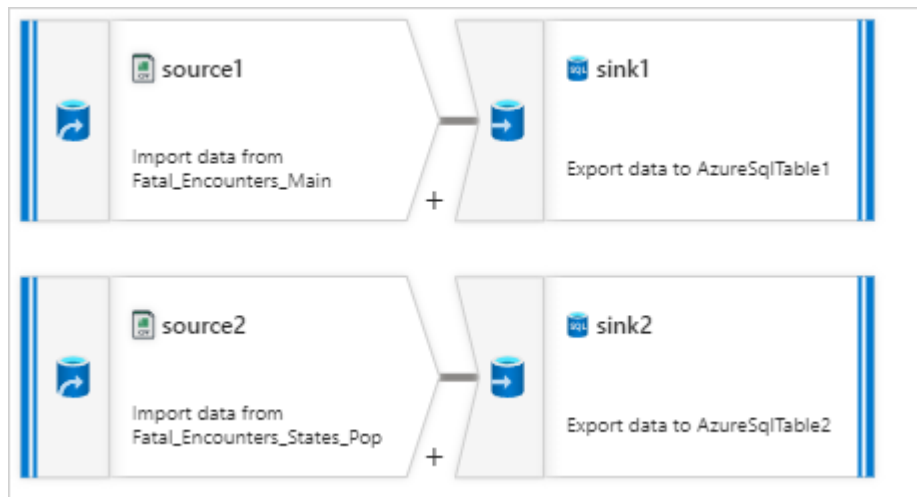
☒ From connection/store

☐ From sample file

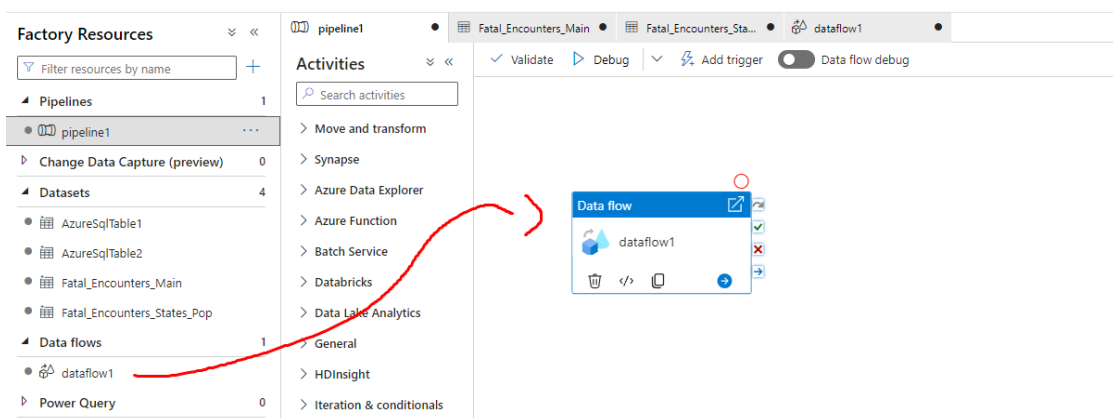
☐ None

The screenshot shows the 'Set properties' dialog for a dataset named 'DelimitedText1'. The 'Name' field is 'DelimitedText1'. The 'Linked service' is 'AzureBlobStorage1'. The 'File path' is 'gabrielcontainer / Directory / fatal\_encounters\_main....'. The 'First row as header' checkbox is checked. The 'Import schema' section has 'From connection/store' selected. The 'Properties' pane on the right shows the 'General' tab with the dataset name and description.

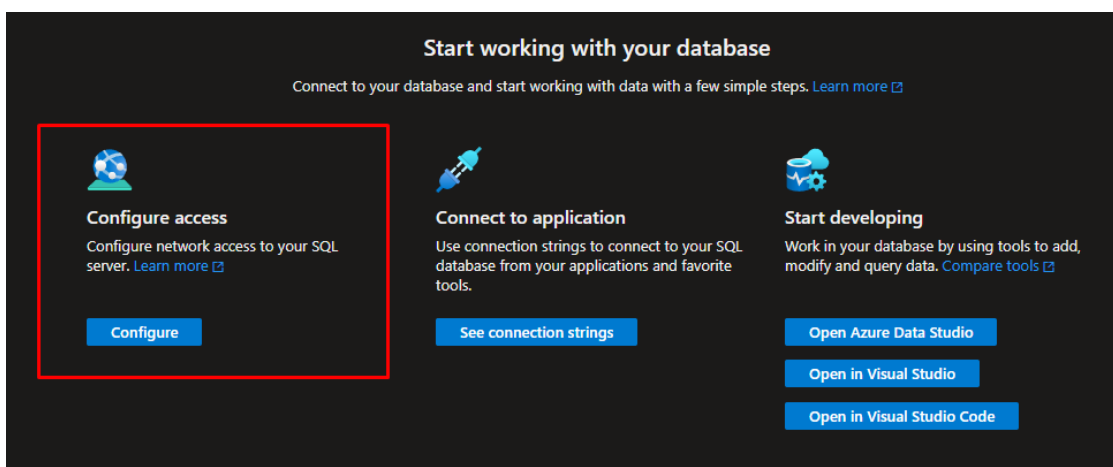
The screenshot shows the 'New linked service' dialog for an 'Azure SQL Database'. The 'Name' field is 'AzureSqlDatabase1'. The 'Description' field is empty. The 'Connect via integration runtime' dropdown is set to 'AutoResolveIntegrationRuntime'. The 'Account selection method' is 'From Azure subscription'. The 'Azure subscription' dropdown is set to 'Azure subscription 1'. The 'Server name' field is 'servername'. The 'Database name' field is 'database'. The 'Authentication type' is 'SQL authentication'. The 'User name' field is 'serveradmin'. The 'Password' field is 'password'. The 'Always encrypted' checkbox is unchecked. The 'Create' button is highlighted.



A imagem acima representa o dataflow, ou seja, não foram feitas alterações nas tabelas. As duas foram, individualmente, exportadas para o servidor do SQL por meio do “sink”. E esse dataflow foi usado para criar o pipeline:



Contudo, para que o processo ocorresse sem erros foi antes necessário configurar o acesso ao servidor SQL:





Após a criação e configuração do servidor SQL, o pipeline pôde ser rodado. Logo em seguida, foi criado um ambiente Databricks:

Home > Azure Databricks >

## Create an Azure Databricks workspace

Basics Networking Encryption Tags Review + create

### Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ Azure subscription 1

Resource group \* ⓘ resourcegroup1  
[Create new](#)

### Instance Details

Workspace name \* gabrielworkspace ✓

Region \* East US ✓

Pricing Tier \* ⓘ Premium (+ Role-based access controls) ✓

Managed Resource Group name Enter name for managed resource group

Foi criado um cluster para poder ser usado no notebook para as análises:

Microsoft Azure databricks Search data, notebooks, recent, and more... CTRL + P

New

- Workspace
- Recents
- Catalog
- Workflows
- Compute

SQL

- SQL Editor
- Queries
- Dashboards
- Alerts
- Query History
- SQL Warehouses

Data Engineering

- Job Runs
- Data Ingestion
- Delta Live Tables

Machine Learning

- Experiments
- Features
- Models
- Serving

Marketplace

- Partner Connect

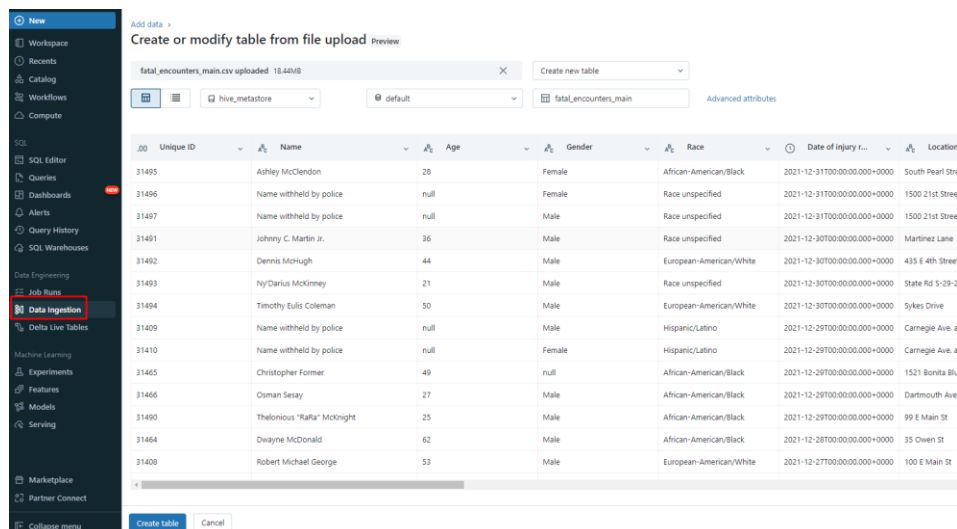
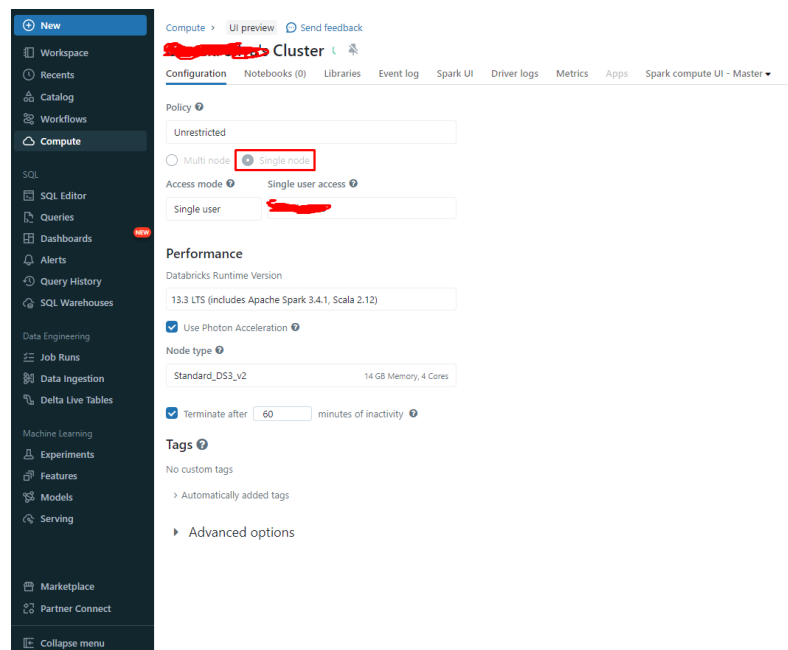
Collapse menu

### Compute

All-purpose compute Job compute SQL warehouses Pools Policies ⓘ

Filter compute you have access to Created by

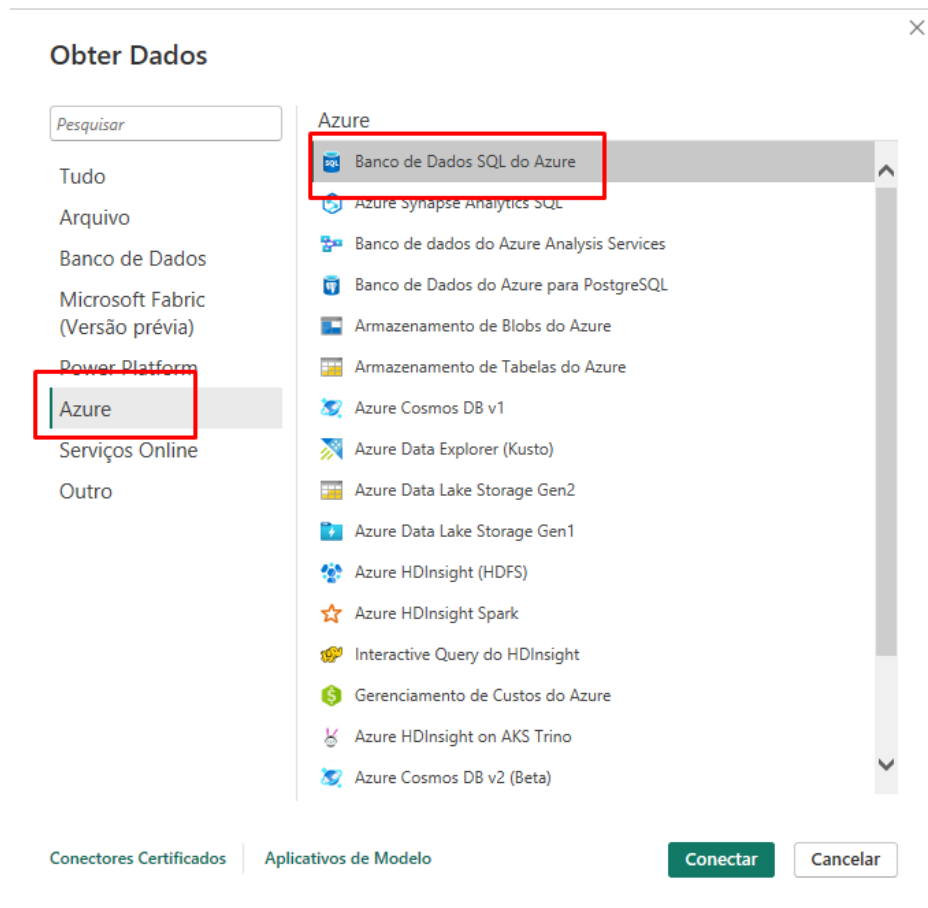
State ⓘ	Name	Policy	Runtime	Active memory	Active cores	Active DB
<div>+</div> <p>No compute</p> <p>Create compute to run workloads from your notebooks and jobs. <a href="#">Learn more about best practices for compute configuration</a></p> <p>Create compute</p>						



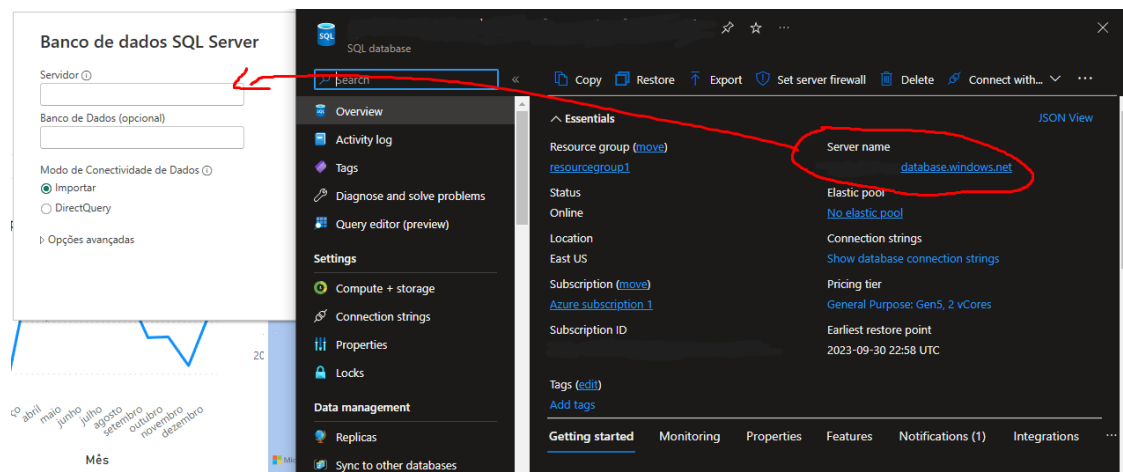
## 4. Power BI

Por meio da base de dados do SQL Server, foi gerado um report interativo no Power BI com os dados de “Fatal Encounters”.

O primeiro passo foi importar os dados. Isso foi feito extraindo-os por meio do Azure SQL Database:



Inserindo a url do banco de dados para extrair:



E, em seguida, selecionando as tabelas:

Navegador

Opções de Exibição

sys.database\_firewall\_rules

fatal\_encounters\_main

fatal\_encounters\_states\_population

tiroteios\_per\_capita\_estados\_ano

fatal\_encounters\_main

Visualização baixada em domingo

Unique ID	Name	Age	Gender	Race
2015.0	Leon Wikoff	26	Male	Race unspecified
2016.0	Mary Onderdonk	75	Female	European-American
20963.0	Melvin Woodyard	45	Male	Race unspecified
2011.0	Jay Vogel	27	Male	Race unspecified
2012.0	Nathan Lee Rossbach	40	Male	Native American/
20962.0	Dwaine Edward Rinesmith	58	Male	Race unspecified
2013.0	Eliche Williams	21	Male	Race unspecified
2014.0	Vernon M. Seals	35	Male	European-American
2009.0	Jeremy D. Abbo	19	Male	Race unspecified
2010.0	Omari S. Davis	19	Male	Race unspecified
2006.0	Arnold L Willets	36	Male	Asian/Pacific Islar
2007.0	Rafael Correa	null	Male	Hispanic/Latino
2008.0	"Juan ""Johnny"" Salazar"	14	Male	Hispanic/Latino
2004.0	Shane Darwin Lynner	38	Male	European-American
2005.0	William Seegert	51	Male	European-American
2003.0	James Thompson	83	Male	Race unspecified
20961.0	Mark Boyce	37	Male	Race unspecified
2001.0	Jesse James Ortiz	25	Male	Race unspecified
2002.0	Stephen Burke	34	Male	Race unspecified
1997.0	Todd Reeves	18	Male	Race unspecified
1998.0	Christopher Talley	24	Male	European-American
1999.0	Ki Yang	46	Male	Asian/Pacific Islar

Selecionar Tabelas Relacionadas

Carregar

Transformar Dados

Cancelar

Na parte de transformação, no editor do Power Query, o tipo das colunas foi alterado para o correto. Além disso, correções pontuais foram feitas nos dados, além de relacionamentos entre tabelas.

O relatório final ficou da seguinte forma:

Fatal Encounters

A step toward creating an impartial, comprehensive and searchable national database of people killed during interactions with police

Legenda:

EF = Encontros Fatais

31,47 Mil

Número total de encontros fatais

Parâmetro (Gráfico de distribuição)

☐ Age
 ☐ Agency or agencies involved
 ☐ Aggressive physical movement
 ☐ Alleged weapon

Ano

Todos

Trimestre

Todos

Mês

Todos

Estado

Todos

Cidade

Todos

Gênero

Todos

Raça

Todos

Agência(s) envolvida(s)

Todos

EF\* por Ano

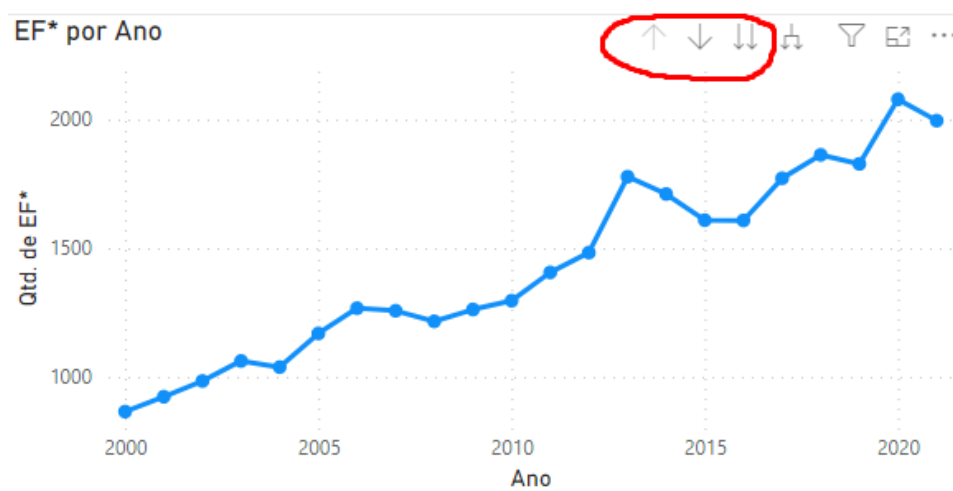
EF\* per capita por Ano

Localização dos EF\*

Qtd. de EF\* por Age

A intenção é deixar com que o usuário explore os dados da maneira que quiser e a maior dificuldade foi, para o gráfico de distribuição (canto inferior direito) fazer com que aparecesse um “Top N” dinamicamente, dado que o eixo X é escolhido pelo filtro “Parâmetro (Gráfico de distribuição)”.

Obs, para “subir” ou “descer” a hierarquia de data, basta usar as setas, isto é, Mês→Trimestre→Ano e vice-versa:



[Link para o relatório](#)

## 5. Análise

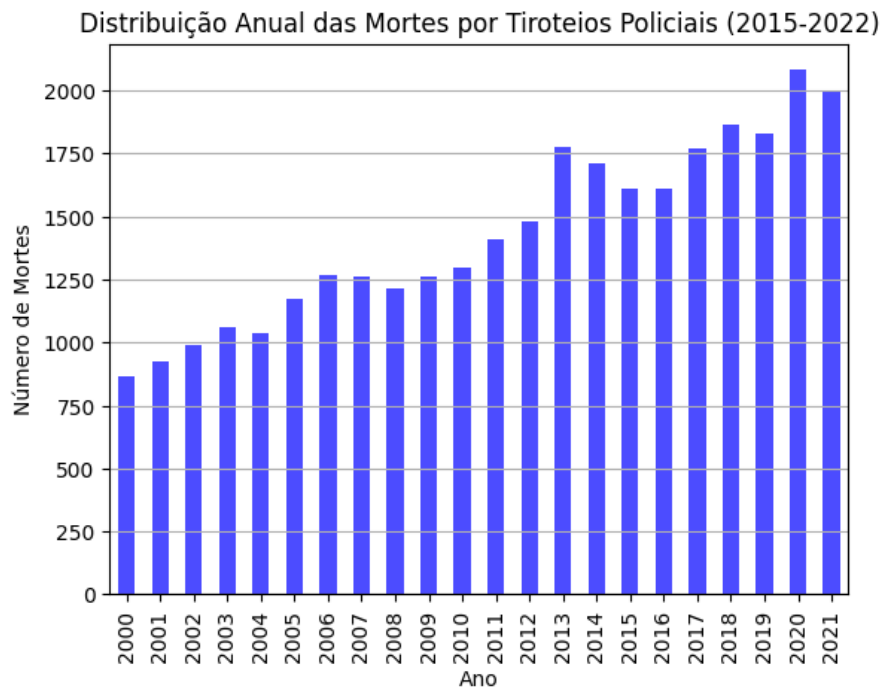
A parte de análise e os objetivos, de maneira geral, estão presente nos arquivos “Análise de Fatal Encounters.ipynb” e “Análise de Fatal Encounters.html” (melhor visualização e interação), mas nesta seção serão detalhados pontos adicionais.

### a. Qualidade dos dados

Algumas colunas possuíam dados inconsistentes, mas como as análises envolviam, individualmente, uma ou poucas colunas, isso foi resolvido e tratado em cada análise como pode ser visto no arquivo de análise.

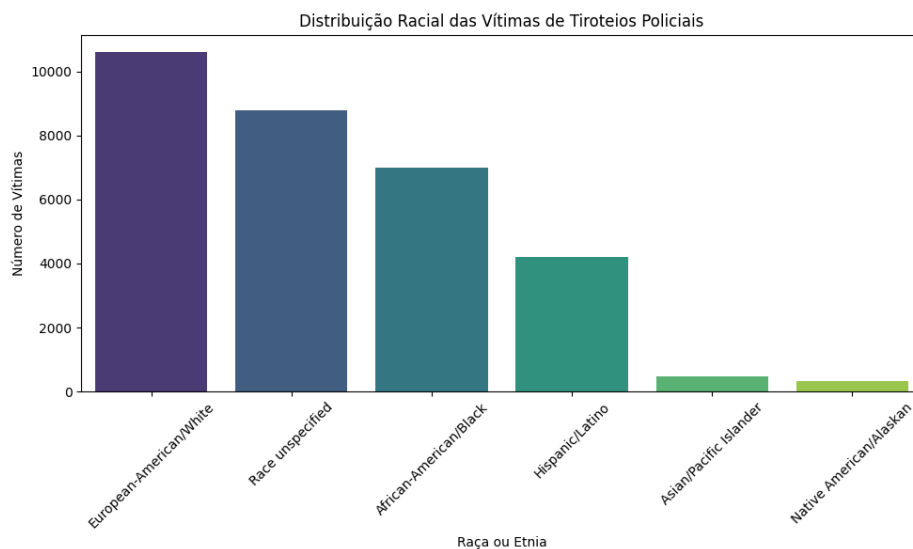
### b. Solução do problema

Qual é a distribuição anual das mortes por tiroteios policiais desde 2015?



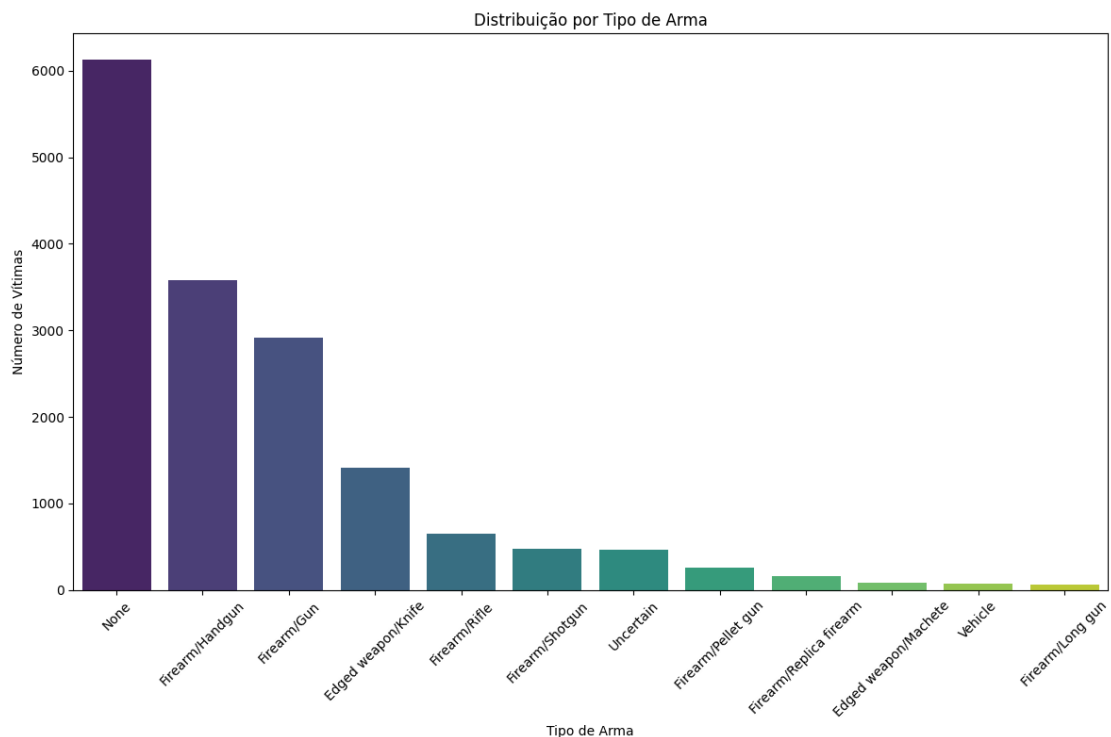
Pode-se observar um aumento no número de mortes por tiroteiros policiais ao longo dos anos, com um pico especialmente em 2013.

Qual é a distribuição racial das vítimas de tiroteios policiais?

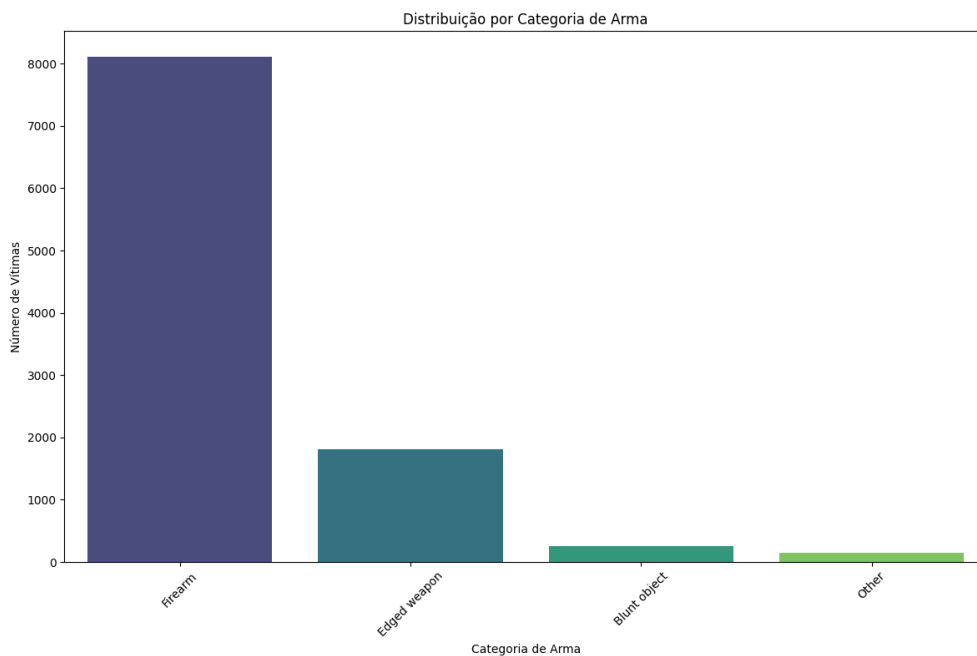


Apesar de “European-American/White” aparecer como a ração com mais casos, a segunda maior é “Race unspecified”, ou seja, essa análise não é conclusiva.

Quantos dos indivíduos mortos estavam armados no momento do tiroteio? E qual era o tipo de arma (se houver)?



Pode-se observar que a maioria dos indivíduos não estava armada, mas, quando estavam, eram, geralmente armas de fogo. Isso pode ser comprovado também pelo gráfico abaixo:

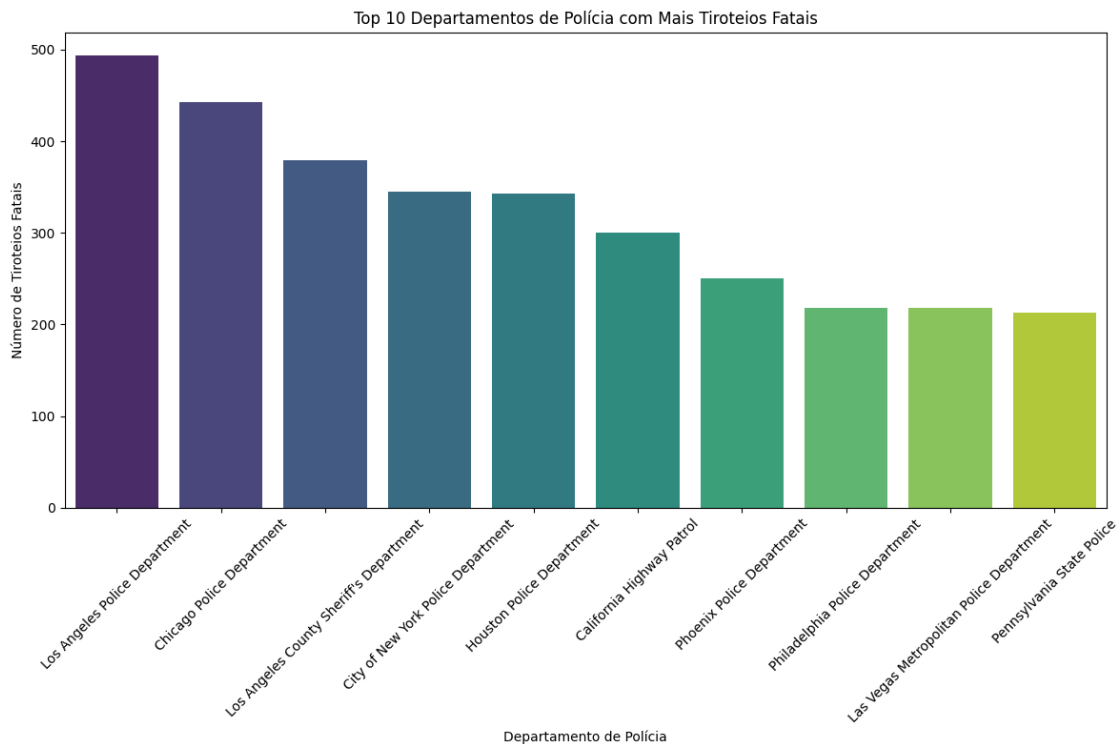


Em quantos dos casos a vítima estava enfrentando uma crise de saúde mental? (Sem resposta)

Essa é uma pergunta interessante, mas por orientação do criador do “Fatal Encounters” essa análise não foi feita. Segue recomendação do autor, presente na

tabela oriunda do site: “Column AC, Were police aware of symptoms of mental illness before interaction? INTERNAL USE, NOT FOR ANALYSIS.”.

Há uma tendência ou padrão nos departamentos de polícia que têm taxas mais altas de tiroteios fatais?



Deve-se ter cuidado nesse tipo de análise por causa de características regionais. Talvez alguns departamentos estejam localizados em áreas com maiores taxas de criminalidade ou outros fatores sociais que podem influenciar as taxas de tiroteios. O uso de uma base de dados adicional pontuando estados/regiões com maiores taxas de criminalidade seria interessante. Obviamente, um departamento de polícia em uma cidade grande naturalmente terá números mais altos do que um em uma cidade pequena, então talvez seja útil normalizar os números (por exemplo, tiroteios per capita). Que é justamente a próxima análise.

Existem regiões ou estados específicos que têm uma incidência desproporcionalmente alta de tiroteios policiais fatais?

Foi criada uma tabela contendo tiroteios/mortes per capita, mas nenhuma análise específica foi feita dado que possui muitos estados e anos. Mas a tabela foi baixada e ingerida no SQL server para ser usada no Power BI.



Cmd 35

```

1 # Essa tabela final permitiria muitas outras análises considerando estados específicos desejados
2 display(df_per_capita_display.sort_values(by=['State', 'Ano']), maxRows=df_per_capita_anos.count())

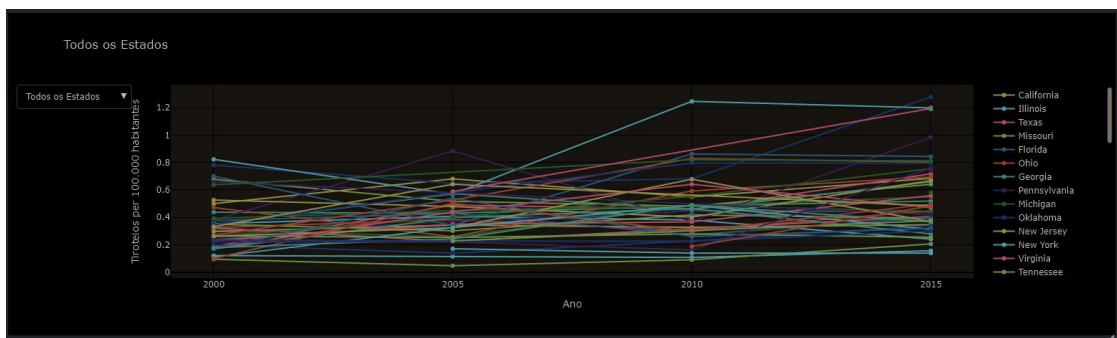
```

Table ▾ +

	State	Count	Population	tiroteios_per_capita	Ano
1	Alabama	12	4452339	0.2695	2000
2	Alabama	24	4557808	0.5266	2005
3	Alabama	40	4822023	0.8295	2010
4	Alabama	39	4858979	0.8026	2015
5	Alaska	4	627500	0.6375	2000
6	Alaska	6	731449	0.8203	2010
7	Alaska	6	738432	0.8125	2015

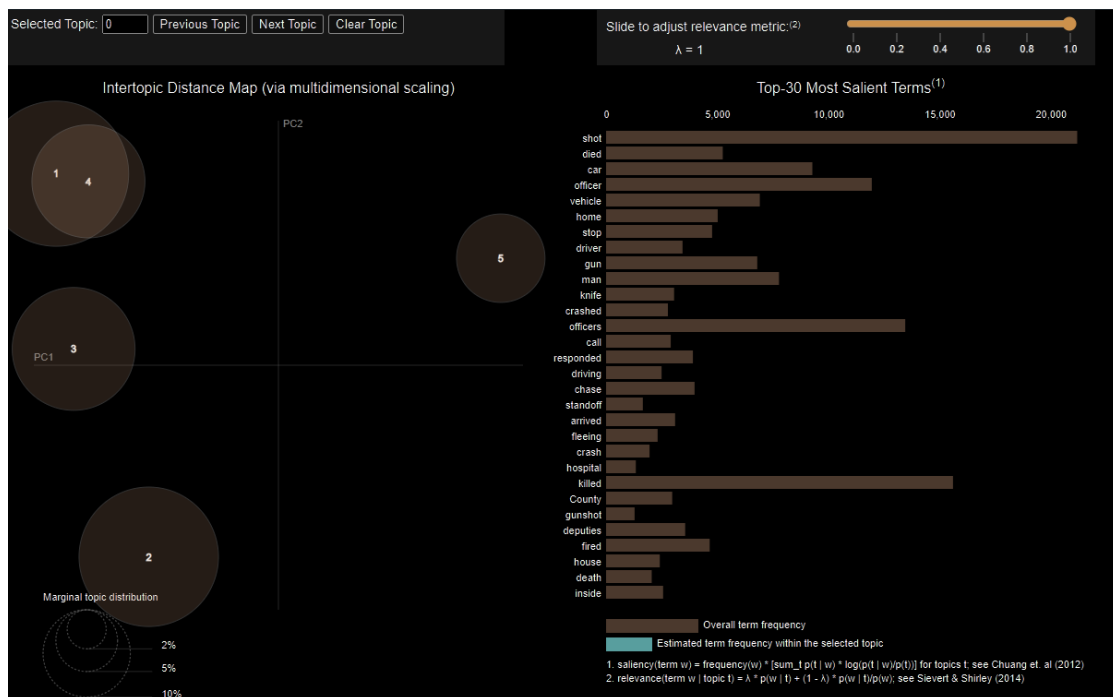
↓ 173 rows | 0.23 seconds runtime

Contudo, no próprio ambiente do notebook, foi criada uma tabela interativa em que era possível selecionar o estado e observar os tiroteios/mortes per capita a cada 5 anos:



Parte de NLP levando em consideração a coluna "Brief Description"

Análise de tópicos usando LDA (Latent Dirichlet Allocation)



Aqui, foi usada a técnica de Latent Dirichlet Allocation (LDA) para identificar tópicos comuns nas descrições. O que poderia revelar padrões ou categorias de incidentes. A visualização com pyLDAvis é interativa e fornece uma ótima forma de explorar os tópicos e as palavras-chave associadas a cada tópico. Isso foi feito de maneira exploratória e para usar segue um detalhamento:

1. Gráfico à esquerda (Espaço bidimensional):

- Cada bolha representa um tópico do modelo LDA.
- A distância entre as bolhas pode ser interpretada como a diferença entre os tópicos. Se as bolhas estiverem mais próximas umas das outras, significa que os tópicos são mais semelhantes.
- O tamanho da bolha reflete a prevalência do tópico na coleção de documentos. Bolhas maiores indicam tópicos que ocorrem mais frequentemente.

2. Barra de deslizamento (slider) " $\lambda$ " (lambda):

- Controla a classificação das palavras-chave exibidas à direita.
- Com  $\lambda = 1$ , as palavras-chave são classificadas apenas por sua probabilidade dentro do tópico.
- Com  $\lambda = 0$ , as palavras-chave são classificadas por sua "exclusividade" ao tópico. Em outras palavras, palavras que são distintas para um

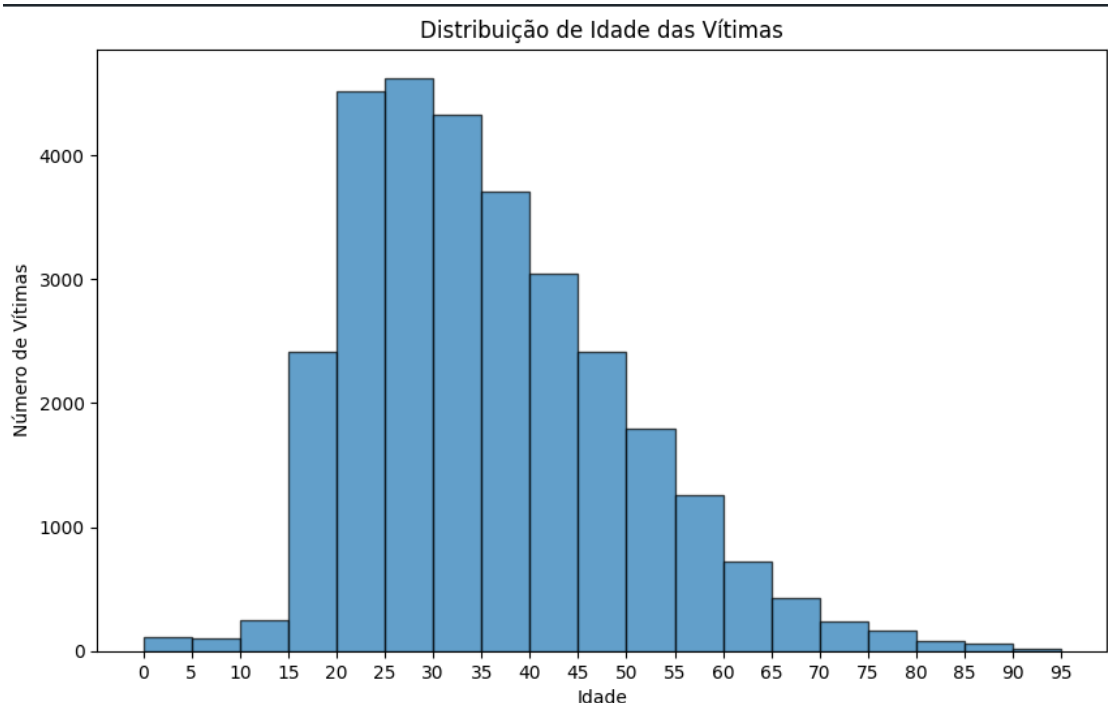
- Ajustar o valor de  $\lambda$  permite que se veja palavras que são mais específicas para um tópico, bem como palavras que são mais frequentes, mas que podem ser compartilhadas por vários tópicos.

- Mostra as palavras-chave mais relevantes para o tópico selecionado.
- O tamanho da barra indica a importância da palavra-chave para o tópico.
- A cor da barra reflete a frequência da palavra-chave em toda a coleção de documentos. Palavras que são comuns em todos os documentos (não apenas no tópico selecionado) são coloridas em vermelho, enquanto palavras que são mais exclusivas para o tópico são coloridas em azul.

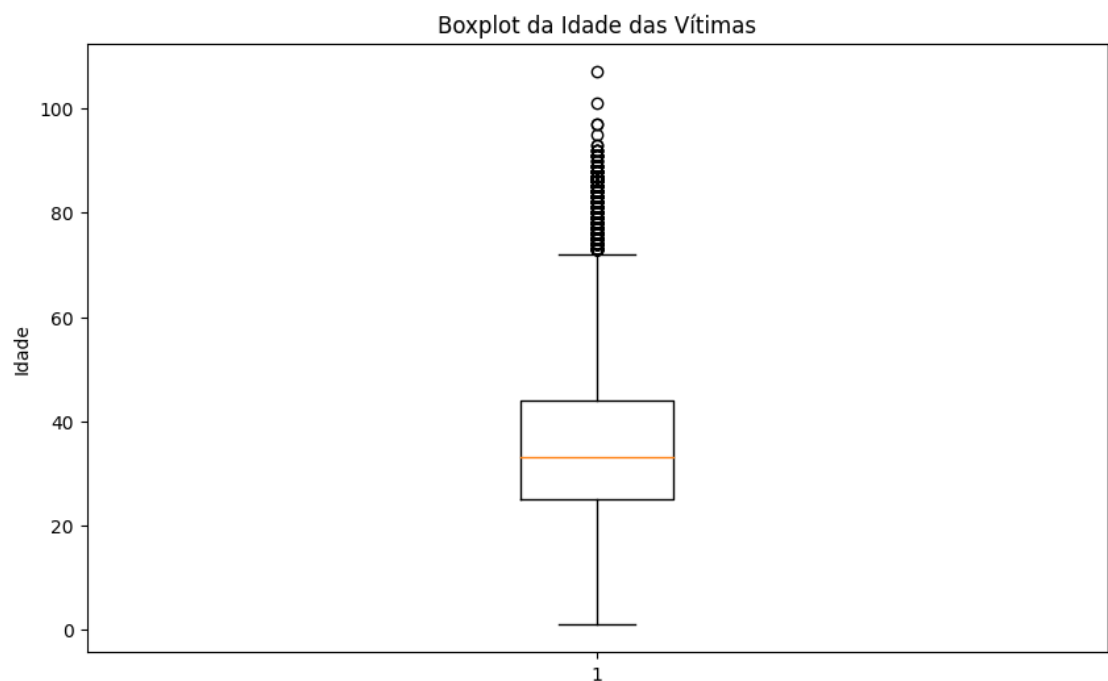
[illegible]

Análise demográfica

Distribuição de idade

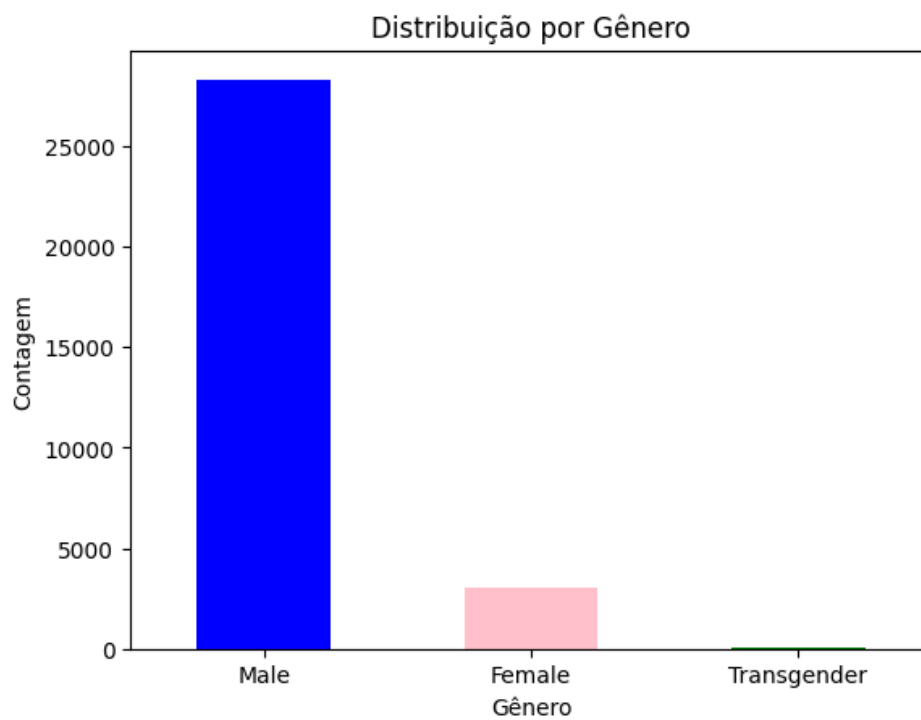


É possível observar que as mortes ocorrem em uma faixa etária “jovem”, entre 20 e 35 anos.



Pelo boxplot, é possível observar que valores acima do limite superior, ou seja, aproximadamente 72 anos, são considerados outliers. Aparentemente, por não possuir uma caixa “grande”, não há alta variabilidade nos dados dentro dos quartis. A mediana parece estar mais próxima ao Q1, então os dados podem ser assimétricos à direita (ou com inclinação positiva).

## Distribuição por gênero



```
Gender
Male      90.251994
Female     9.668262
Transgender 0.073365
Name: proportion, dtype: float64
```

Majoritariamente, os casos ocorrem em homens, que representam mais de 90% das vítimas.

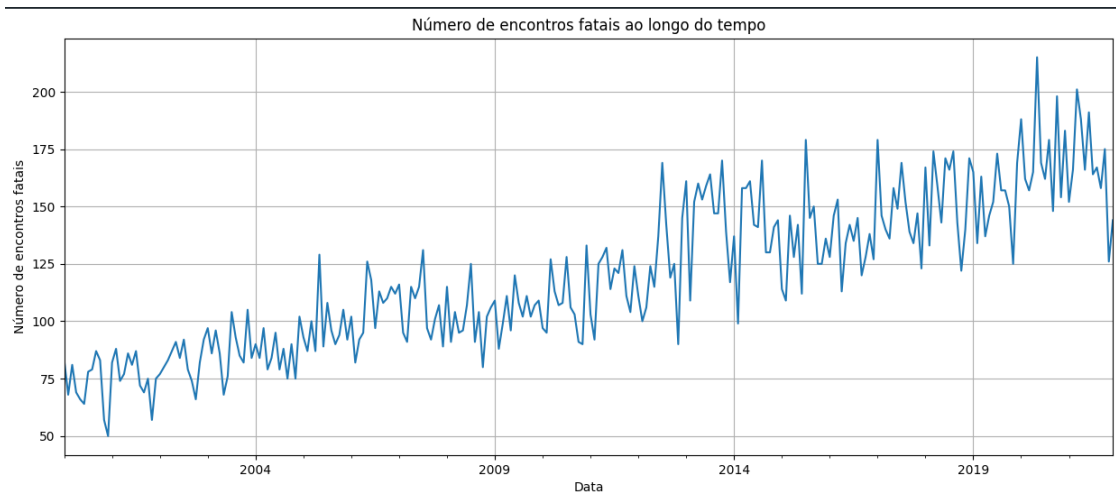
## Distribuição por raça

Para fazer uma comparação com a População Geral, seria necessário ter os dados da distribuição racial da população geral.

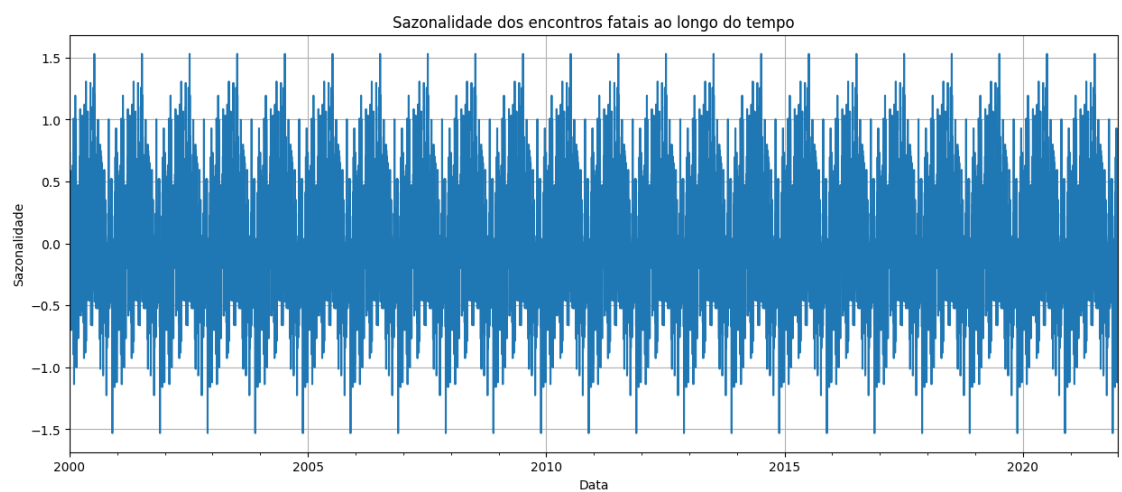
## Análise temporal

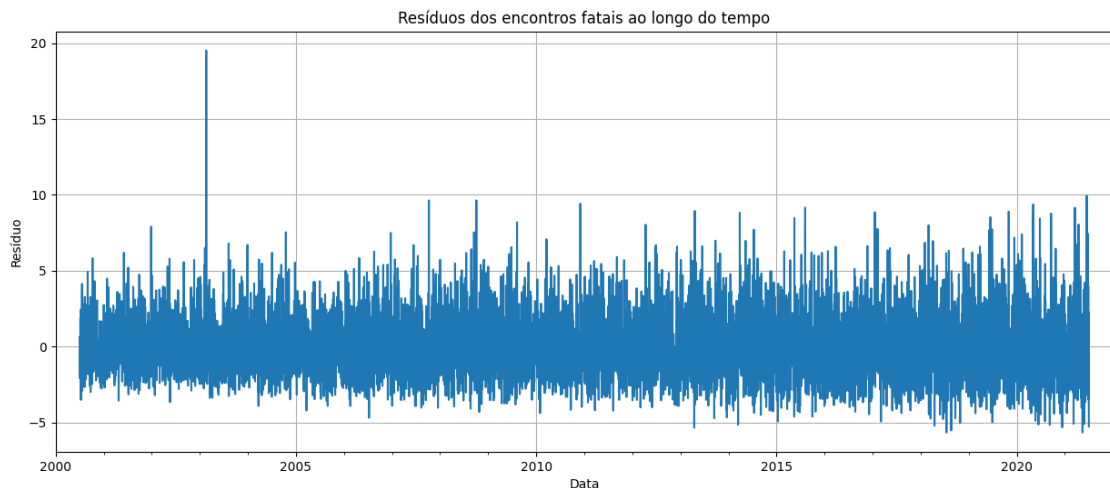
Número de incidentes fatais ao longo do tempo (anualmente, mensalmente, etc.).

Identificar se há algum padrão sazonal nos encontros fatais.



Primeiramente, foi feita uma análise do número de encontros fatais. Nesse gráfico, cada ponto representa um mês de um ano específico. É possível observar uma tendência de aumento ao longo dos anos.





Em seguida, foi feita uma decomposição sazonal, uma técnica que permite dividir uma série temporal em seus componentes: tendência, sazonalidade e resíduo.

#### Componente de Tendência:

No gráfico da tendência, pode ser observada a evolução geral dos encontros fatais ao longo do tempo, sem considerar flutuações sazonais.

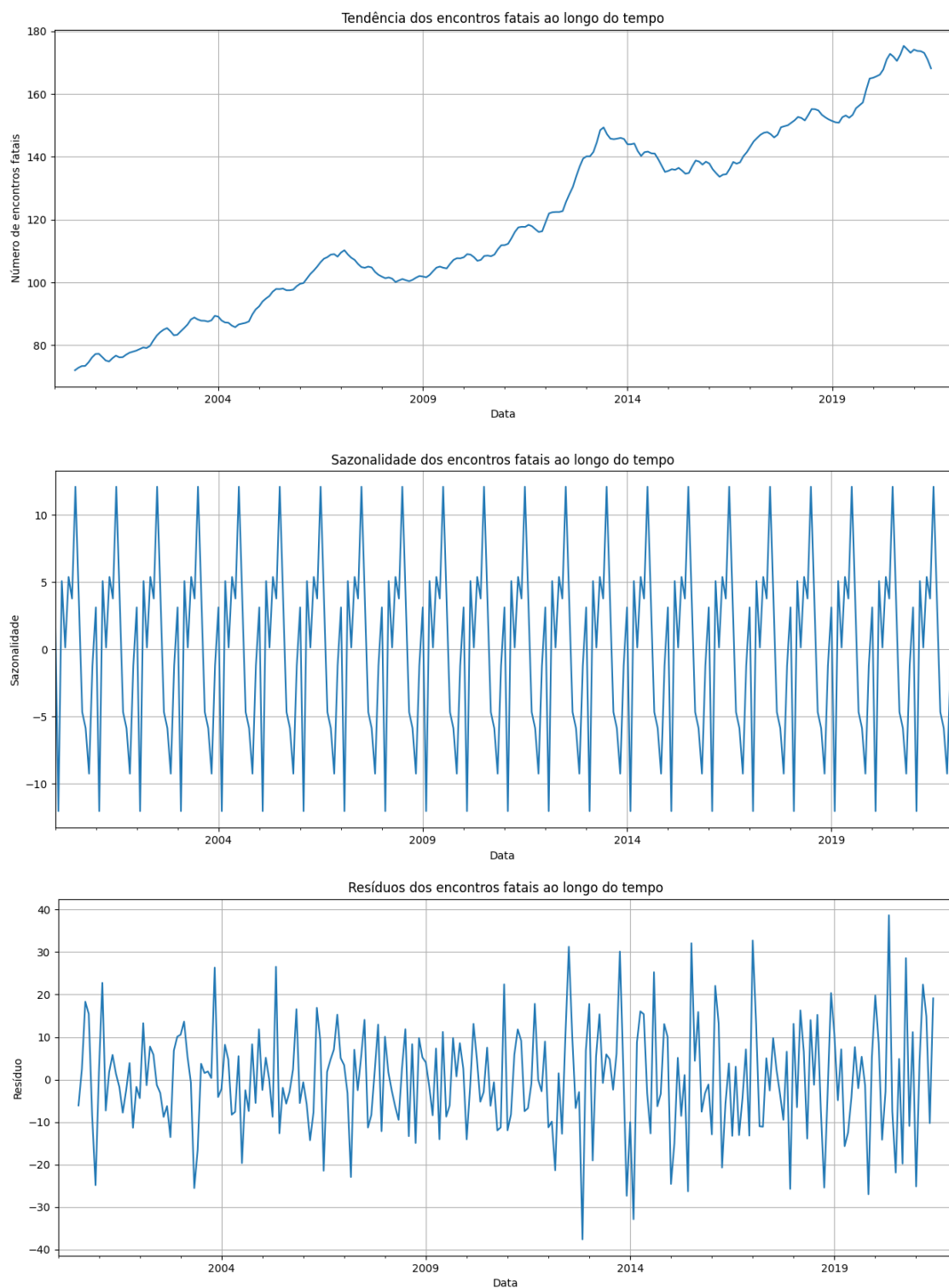
#### Componente de Sazonalidade:

Neste gráfico, é possível observar os padrões que se repetem em intervalos regulares. No caso dos encontros fatais, pode haver picos e vales. Se o valor no eixo Y for -1,5, isso pode indicar que, em média, os encontros fatais, nesse período específico, são 1,5 unidades abaixo da média anual. Em contraste, se o valor for +1,5, isso indica que os encontros fatais são, em média, 1,5 unidades acima da média anual.

#### Componente Residual:

O resíduo mostra as flutuações que não podem ser atribuídas nem à tendência nem à sazonalidade. São as "anomalias" ou "ruídos" na série. Se, por exemplo, houve algum evento que afetou os encontros fatais e que não é uma característica recorrente, isso aparecerá no gráfico residual.

Essa análise foi feita novamente, mas considerando meses, para melhor visualização:



Comparar a evolução dos incidentes em diferentes estados ou cidades ao longo dos anos.

Presente no Power BI



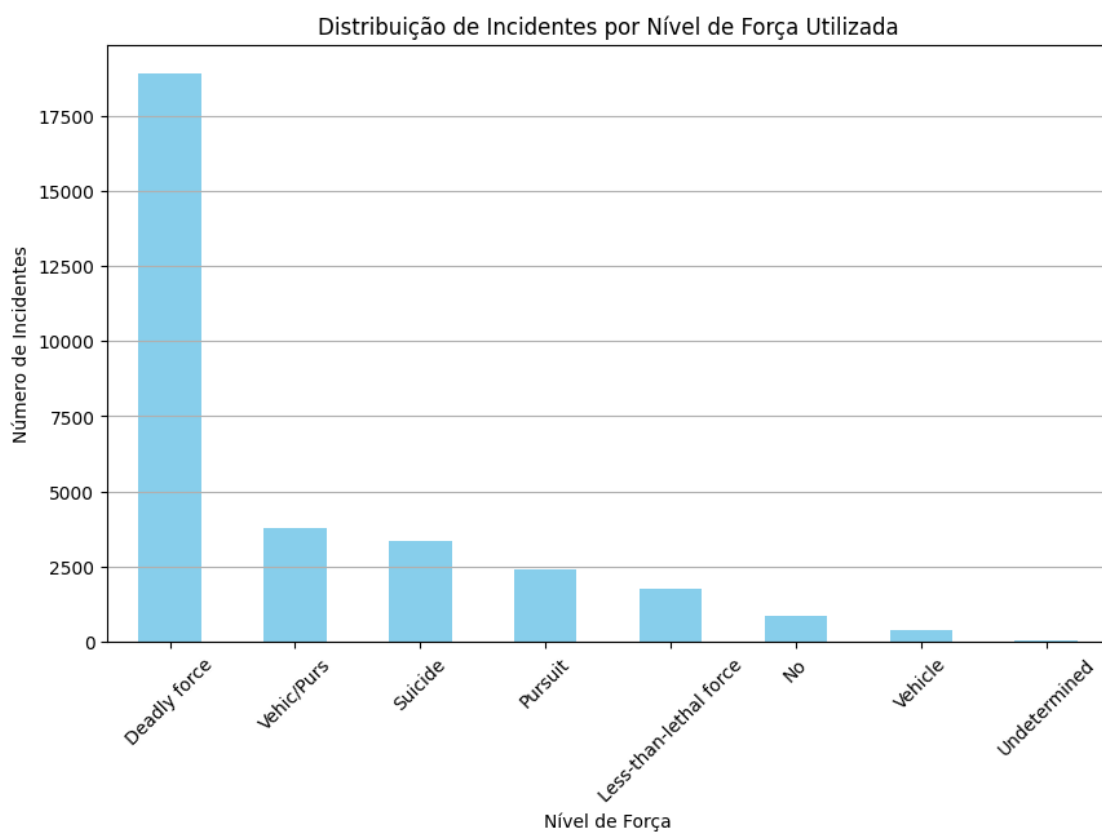
Análise geográfica

Mapa interativo marcando os locais de todos os incidentes fatais.

Presente no Power BI

Análise das circunstâncias

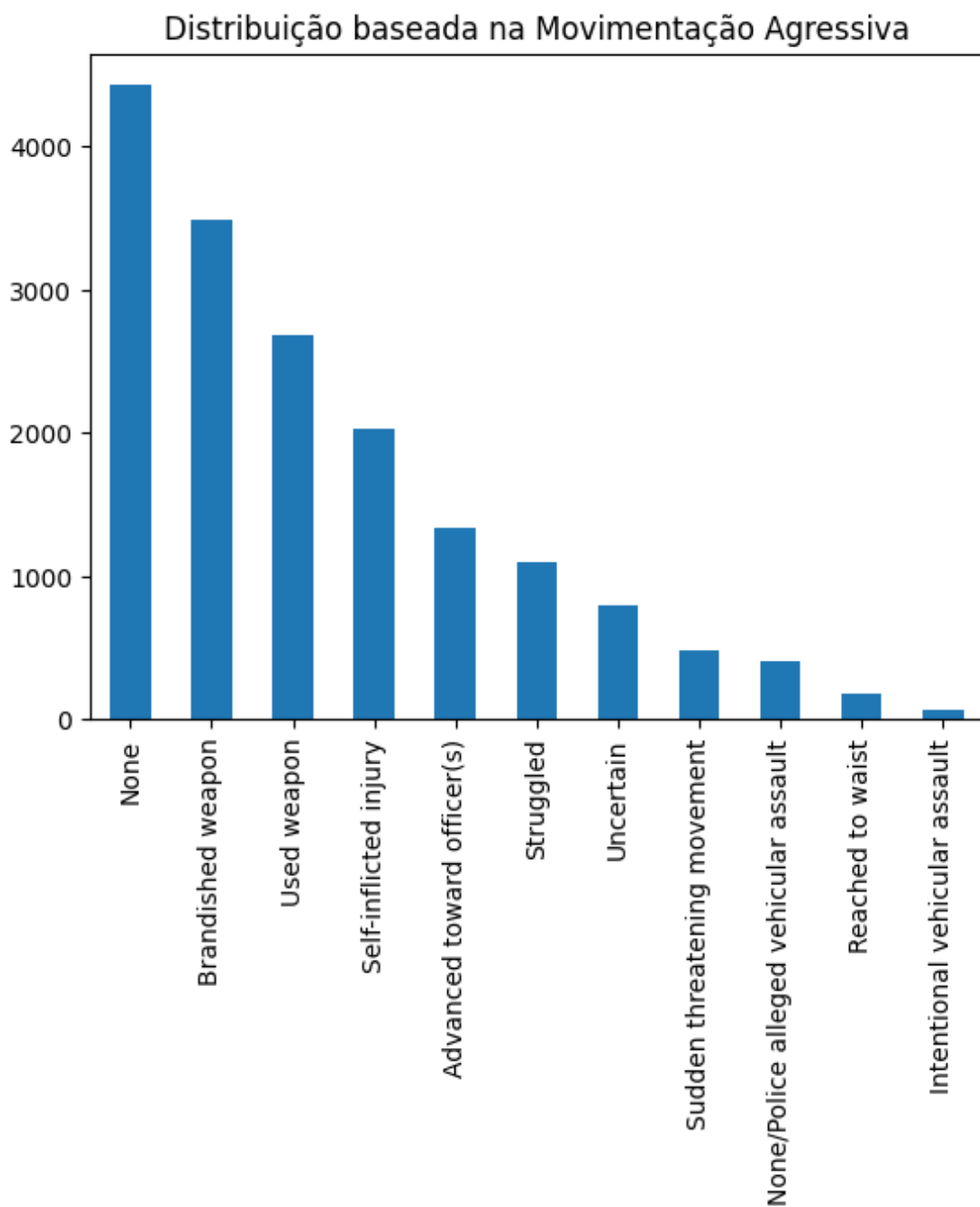
Distribuição de incidentes com base no nível mais alto de força utilizada.



A descrição dessa coluna na tabela original é confusa, mas, aparentemente, é o maior nível de força utilizado pela polícia. Com isso, é possível observar que, na maior parte dos casos, houve a intenção de matar com uso força mortal.

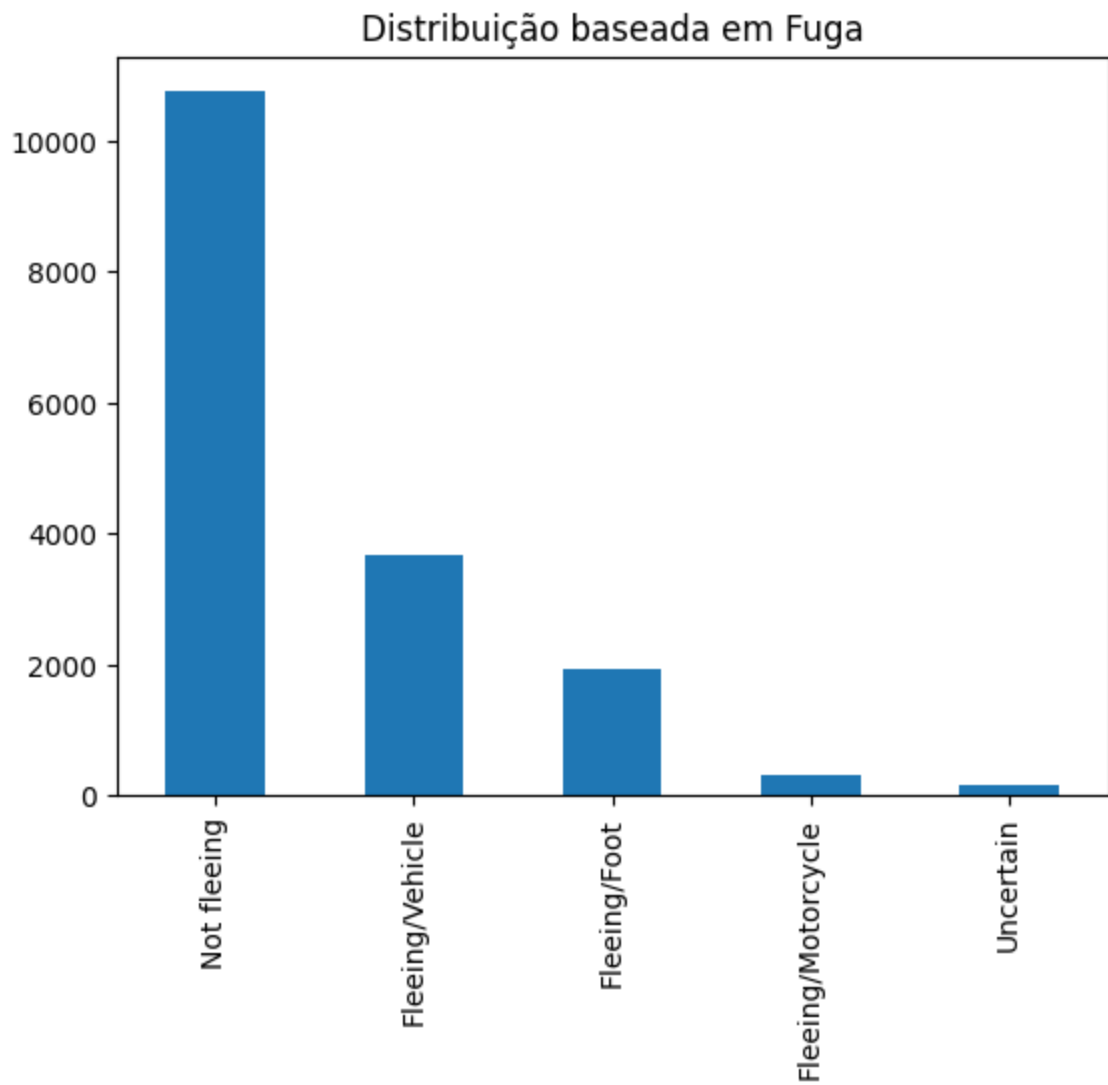
Análise de incidentes envolvendo fuga ou movimentação agressiva.

1. Distribuição de incidentes baseada na movimentação agressiva



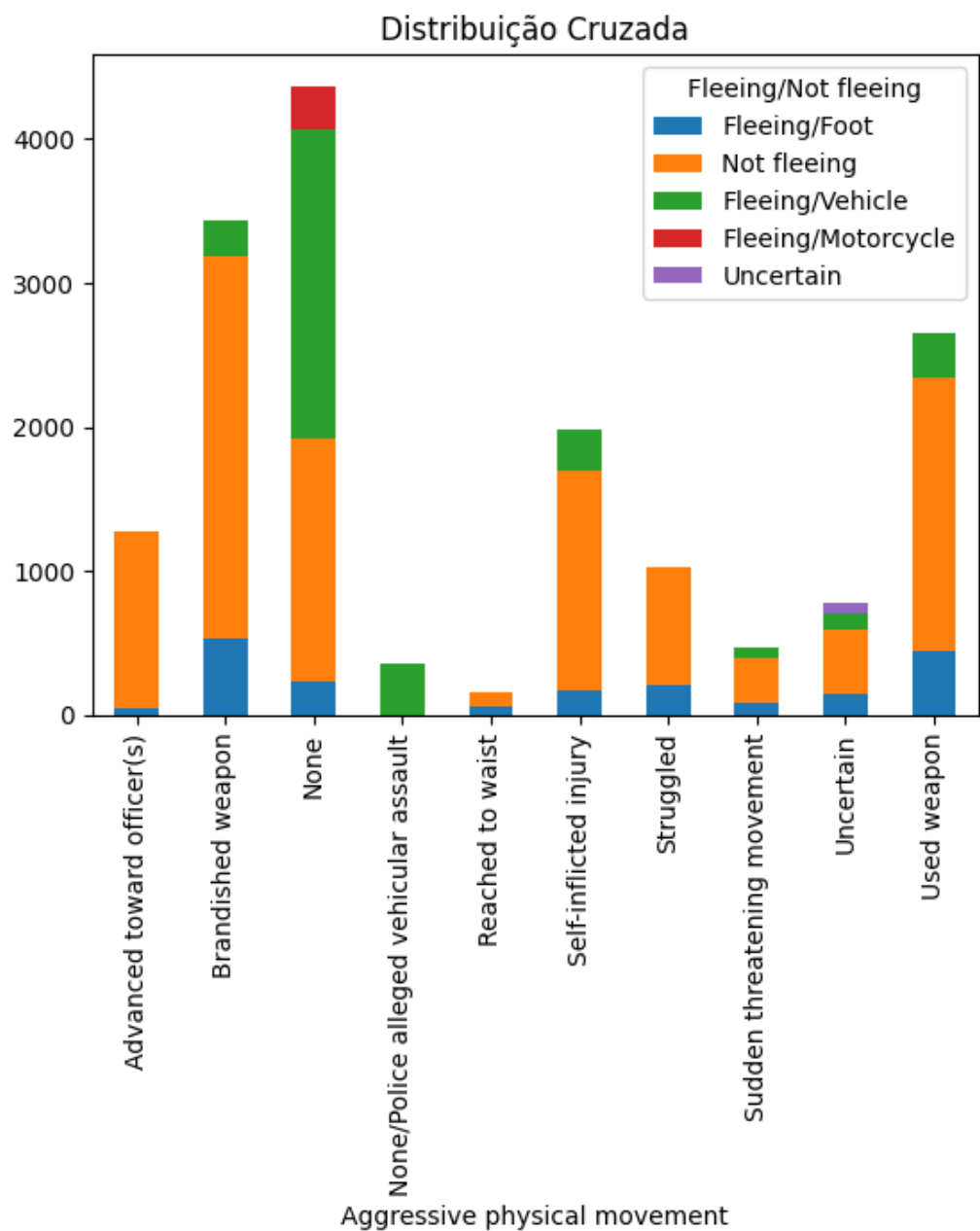
A maioria dos indivíduos não apresentou movimentação agressiva.

## 2. Distribuição de incidentes baseada em fuga



A maioria dos indivíduos não estava fugindo.

### 3. Cruzando as duas colunas



#### Análises Comparativas:

Comparar os encontros fatais de diferentes estados ou cidades considerando variáveis socioeconômicas, como renda média, nível de educação e taxas de criminalidade.

Investigar correlações entre encontros fatais e outros fatores sociodemográficos.

De maneira geral, os objetivos do trabalho foram respondidos, mas seriam interessantes bases adicionais para enriquecer as análises.