



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. fatalencounters.org</p> <p>To download, go to the Google spreadsheet, link below, go under File>Download as> and pick your format. We recommend comma-separated values.</p> <p>https://docs.google.com/spreadsheets/d/1dKmaV_JMwG8XBzRyP8b4d9C0pka7L7mvsyzvAoE/edit#gid=0</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 d.burghart@fatalencounters.org. 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 *

Resource group *

[Create new](#)

Instance details

Storage account name *

Region *
[Deploy to an edge zone](#)

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

Redundancy *

[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

Essentials

Resource group: resourcegroup1

Location: East US

Primary/Secondary Location: Primary: East US, Secondary: West US

Subscription: Azure subscription 1

Subscription ID: [redacted]

Disk state: Primary: Available, Secondary: Available

Tags: Add tags

Properties

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: Disabled

Storage account key access: Disabled

Minimum TLS version: Disabled

Infrastructure encryption: Disabled

Networking

Allow access from: Disabled

Number of private endpoint connections: Disabled

Network routing: Disabled

Access for trusted Microsoft services: Disabled

Endpoint type: Disabled

Upload blob

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

Drag and drop files here or [Browse for files](#)

Select an existing container:

[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:

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 *

servername-xxxxxx

.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

Basics

Networking

Security

Additional settings

Tags

Review + 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 *

Azure subscription 1

Resource group *

resourcegroup1

Create new

Database details

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

Database name *

(new)

Server *

(new) (East US)

Create new

Want to use SQL elastic pool?

☐ Yes

☒ No

Workload environment

☐ Development

☒ Production

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

Review + create

Next: Networking >

Cost summary

General Purpose (GP_Gen5_2)

Cost per vCore (in USD)

184.09

x 2

Cost per GB (in USD)

0.12

Max storage selected (in GB)

x 41.6

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 sub-header 'Project details' is followed by the instruction: 'Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.' The 'Subscription' dropdown is set to 'Azure subscription 1'. The 'Resource group' dropdown is set to 'resourcegroup1', with a 'Create new' link below it. The 'Instance details' section includes: 'Name' set to 'gabrielfactory', 'Region' set to 'East US', and 'Version' set to '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' configuration window 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' and shows configuration for 'Azure Blob Storage'. Fields include: 'Name' (AzureBlobStorage1), 'Description', 'Connect via integration runtime' (AutoResolveIntegrationRuntime), 'Authentication type' (Account key), 'Account selection method' (From Azure subscription), 'Azure subscription' (selected), and 'Storage account name' (selected). There are sections for 'Additional connection properties', 'Test connection' (radio buttons for 'To linked service' and 'To file path'), 'Annotations', 'Parameters', and 'Advanced'. 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

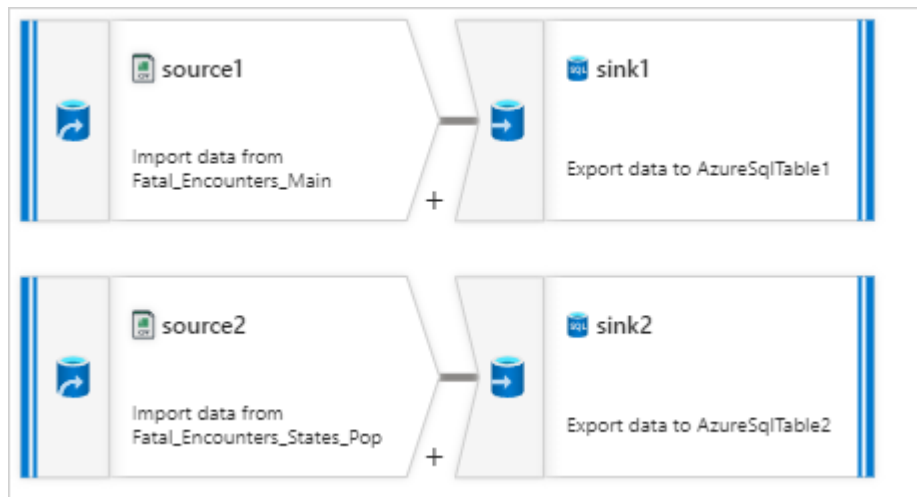
☒ From connection/store

☐ From sample file

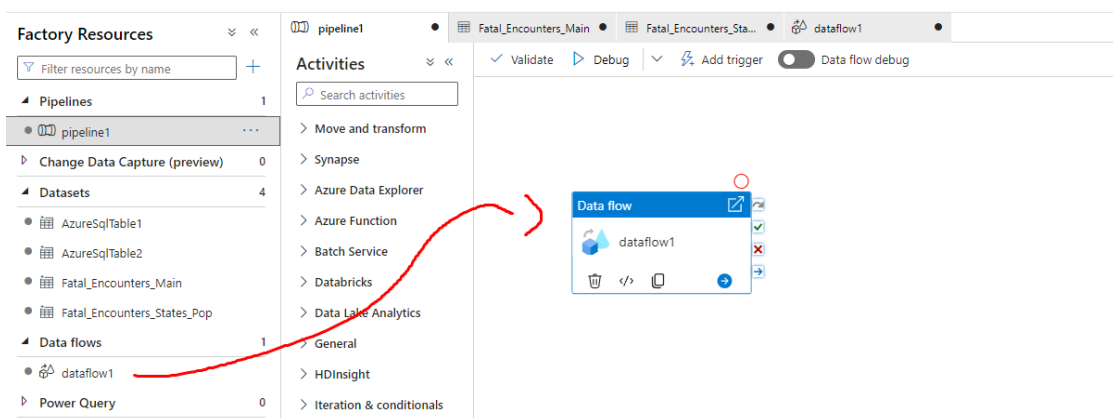
☐ None

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.

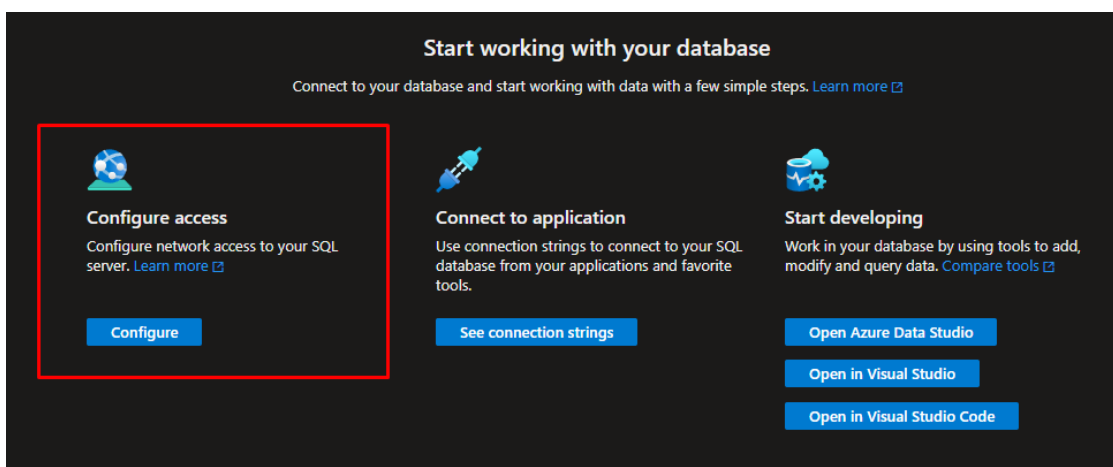
The screenshot shows the 'New linked service' dialog for an 'Azure SQL Database'. The 'Name' field is 'AzureSqlDatabase1'. The 'Description' field is empty. The 'Connect via integration runtime' dropdown is set to 'AutoResolveIntegrationRuntime'. The 'Account selection method' is 'From Azure subscription'. 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

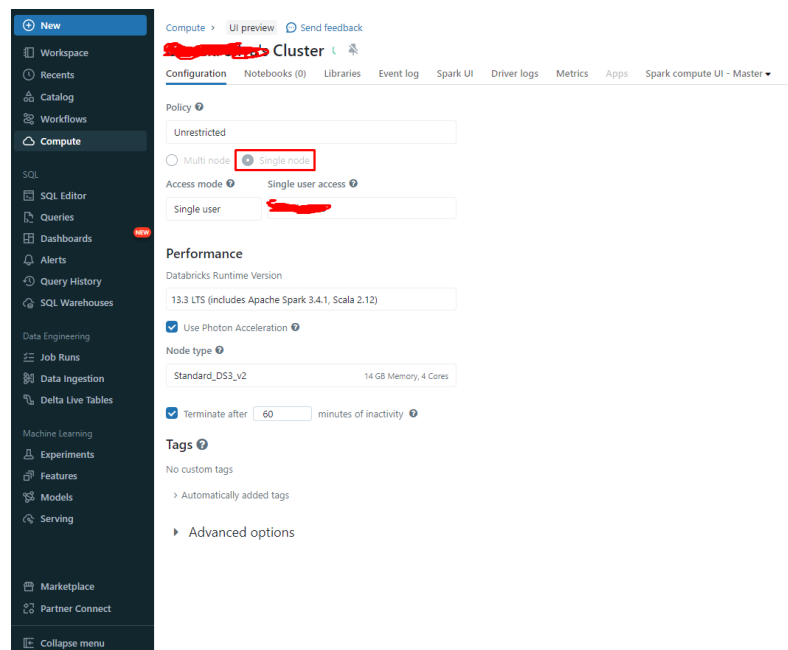
Collapse menu

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. Learn more about best practices for compute configuration</p> <div>Create compute</div>						



New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Engineering

Data Ingestion

Delta Live Tables

Machine Learning

Experiments

Features

Models

Serving

Marketplace

Partner Connect

Collaborate menu

Add data >

Create or modify table from file upload Preview

fatal_encounters_main.csv uploaded 18.44MB

Create new table

hive_metastore

default

fatal_encounters_main

Advanced attributes

ID	Unique ID	Name	Age	Gender	Race	Date of injury	Location
31495	Ashley McClendon	28	Female	African-American/Black	2021-12-31T00:00:00.000+0000	South Pearl Street	
31496	Name withheld by police	null	Female	Race unspecified	2021-12-31T00:00:00.000+0000	1500 21st Street	
31497	Name withheld by police	null	Male	Race unspecified	2021-12-31T00:00:00.000+0000	1500 21st Street	
31491	Johnny C. Martin Jr.	36	Male	Race unspecified	2021-12-30T00:00:00.000+0000	Martinez Lane	
31492	Dennis McKHugh	44	Male	European-American/White	2021-12-30T00:00:00.000+0000	435 E 4th Street	
31493	Ny'Darius McKinney	21	Male	Race unspecified	2021-12-30T00:00:00.000+0000	State Rd 5-29-21	
31494	Timothy Ellis Coleman	50	Male	European-American/White	2021-12-30T00:00:00.000+0000	Sykes Drive	
31409	Name withheld by police	null	Male	Hispanic/Latino	2021-12-29T00:00:00.000+0000	Carnegie Ave. at 1st	
31410	Name withheld by police	null	Female	Hispanic/Latino	2021-12-29T00:00:00.000+0000	Carnegie Ave. at 1st	
31465	Christopher Forner	49	null	African-American/Black	2021-12-29T00:00:00.000+0000	1521 Bonita Blvd	
31466	Oman Sesay	27	Male	African-American/Black	2021-12-29T00:00:00.000+0000	Dartmouth Ave	
31490	Thelonious "Rafra" McInight	25	Male	African-American/Black	2021-12-29T00:00:00.000+0000	99 E Main St	
31464	Dwayne McDonald	62	Male	African-American/Black	2021-12-28T00:00:00.000+0000	35 Owen St	
31408	Robert Michael George	53	Male	European-American/White	2021-12-27T00:00:00.000+0000	100 E Main St	

Create table

Cancel

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”.

[Link para o relatório](#)

5. Análise

A parte de análise, de maneira geral, está presente no arquivo “Análise de Fatal Encounters.ipynb”, 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

Os problemas, de maneira geral, foram resolvidos e podem ser vistos no arquivo de análise.