MVP 03 – Engenharia de Dados

Trabalho desenvolvido por Eduardo Ferreira Gonçalves

Setembro Amarelo - Prevenção ao Suicídio

I. Introdução

Considerando que tenho me dedicado à aplicação da ciência de dados à saúde e aproveitando que o desenvolvimento e a entrega deste MVP serão em setembro, mês dedicado à conscientização sobre a prevenção do suicídio, decidiu-se fazer um estudo para traçar o perfil das pessoas que tiraram a própria vida no intervalo dos anos de 2019 a 2022.

O suicídio é um sério problema de saúde pública que, frequentemente, pode ser prevenido. Embora o tema seja sensível, é crucial abordá-lo e apresentar as medidas preventivas. Muitas pessoas podem acreditar que o suicídio seja uma realidade distante que afeta poucas pessoas, mas, lamentavelmente, as estatísticas da Organização Mundial de Saúde (OMS) revelam o contrário. Segundo a OMS, a cada 40 segundos, uma pessoa tira a própria vida em algum lugar do mundo. Isso se traduz em mais de 703 mil vidas perdidas dessa maneira a cada ano.

II. Objetivo

Conforme mencionado anteriormente, o objetivo deste estudo é a identificação e a análise do perfil das pessoas que cometem suicídio, com a finalidade de identificar características particulares ou padrões que possam ajudar na análise da situação deste problema de saúde pública, inclusive com informação de prevalência. Coletando evidências para ajudar no planejamento e o desenvolvimento de políticas públicas voltadas para sua prevenção.

III. Escolha da plataforma de computação em nuvem

A escolha da Google Cloud Plataform – GCP para fazer o pipeline dos dados foi consequência da ponderação das seguintes características:

1. Grande quantidade de serviços disponibilizados de forma gratuita, mesmo com limite de dados processados ou trafegados;
2. Para novos usuários, a plataforma oferece US$ 300,00 de crédito para usar em 90 dias;
3. As ferramentas do Google têm um padrão de usabilidade que gera uma familiaridade, mesmo nunca tendo operado algum dos serviços disponibilizados;
4. A principal ferramenta de armazenamento e análise é o BigQuery, uma espécie de banco de dados colunar que os apresenta na forma de tabela com semântica de transação similar ao SQL;
5. Integração facilitada do BigQuery com o Google Colab, ferramenta útil para fazer a análise exploratória final dos dados;

Disto exposto, foi criada uma conta nova no Gmail para aproveitar o crédito promocional.

IV. Busca dos dados

Diante da proposta apresentada, iniciou-se a busca pelos dados que pudessem possibilitar a geração de um perfil mais completo possível.

Em apertada síntese, depois de alguns e-mails enviados ao Ministério da Saúde solicitando a informação da existência de uma API que disponibilizasse dados, encontramos a página do site do Departamento de Informática do Sistema Único de Saúde (DATASUS) destinado à Transferência de Arquivos https://datasus.saude.gov.br/transferencia-de-arquivos.

Inicialmente destaca-se que, embora o formulário para download informe que os relatórios de Declarações de Óbitos por causas externas seja até 2021, podemos observar que está sendo atualizado e disponibilizado até o presente ano.

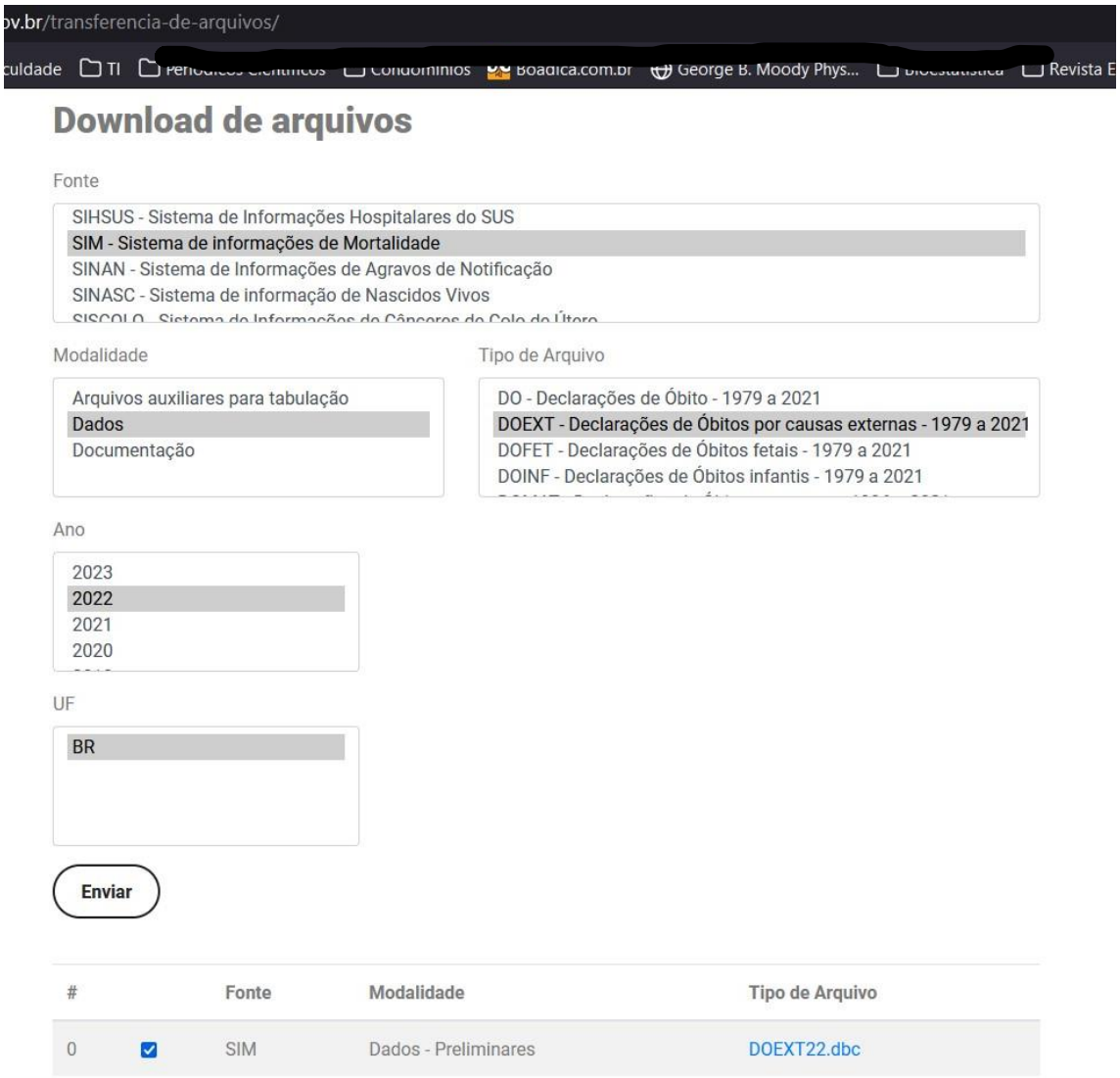


Figura 1: Site DATASUS

V. Licença dos dados

No site do DATASUS não se encontrou documentação acerca da licença de uso dos dados disponibilizados, contudo no capítulo 8 do Guia de dados abertos do Núcleo De Informação e Coordenação do Ponto Br – NIC.br é possível verificar que os dados disponibilizados pelo governo deveriam ser de sob uma licença aberta, justificando que “uma ‘licença aberta’ é importante também para que a Lei de Acesso à Informação seja respeitada”, entretanto, apontam que isso não é um ponto pacífico.

> A Infraestrutura Nacional de Dados Abertos, do governo federal, ainda não chegou a uma conclusão se o ordenamento jurídico brasileiro é suficiente para tratar os dados governamentais de forma aberta, sem a necessidade de licenças, se outras licenças que já existem se

enquadram no contexto brasileiro, ou se será necessário criar uma licença específica para o Brasil.

Cumpre destacar que, conforme o Guia de dados abertos, a licença aberta dá a autorização para que esses dados sejam utilizados por qualquer pessoa, sem restrições, e para qualquer fim, inclusive comerciais.

VI. Coleta dos dados

Quanto à fonte de dados, deve-se pontuar que alguns problemas foram identificados:

O primeiro é que o download não funciona com o botão de downloads, o link é um endereço *ftp* e desde 2021 os navegadores deixaram de dar suporte ao download com links de *ftp*. Consequentemente, foi necessário obter o endereço de cada conjunto de dados para fazer download de outras formas. A saída encontrada foi através do comando *wget* do Linux.

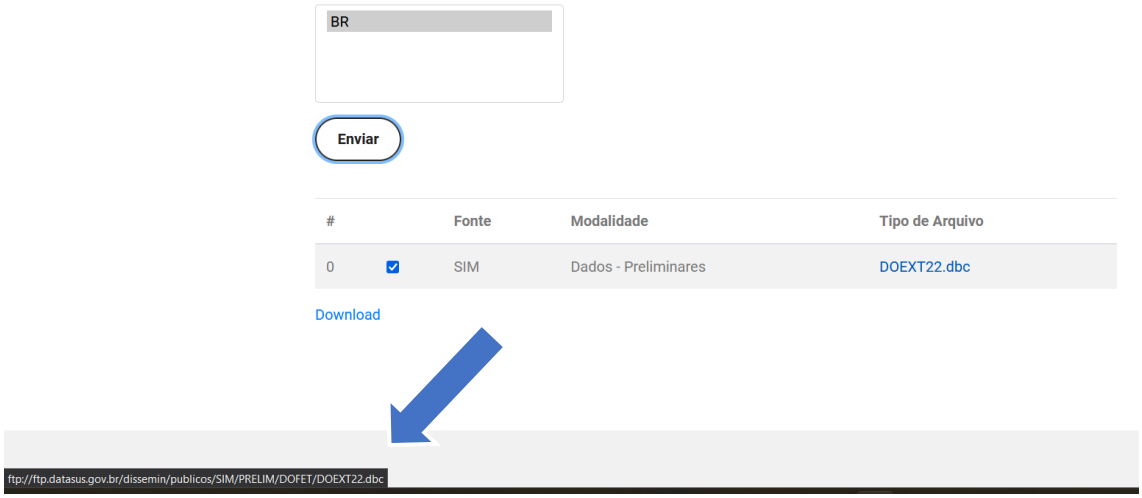


Figura 2: link em ftp

Aproveitando que o ambiente disponibilizado pelo GCP possui um ambiente Linux chamado Cloud Shell e um editor de código similar ao VS Code integrado, foram feitos scripts em Python para a automatização de alguns procedimentos.

Um procedimento automatizado foi o download dos arquivos do DATASUS e sua movimentação para uma pasta específica.

```python
import os

#criação da pasta para salvar os dados
try:
    os.mkdir('base_dados_dbc')

except FileExistsError:
    pass


# download da base de dados do ano de 2022
url = "ftp://ftp.datasus.gov.br/dissemin/publicos/SIM/PRELIM/DOFET/DOEXT22.dbc"
os.system(f'wget {url}')


#download das bases de dados dos anos de 2021 até 2019
anos = ['21', '20', '19']
for ano in anos:
    os.system(f'wget ftp://ftp.datasus.gov.br/dissemin/publicos/SIM/CID10/DOFET/DOEXT{ano}.dbc')


# mover os dados para a pasta criada
anos = ['22', '21', '20', '19']
for ano in anos:
    os.system(f'mv DOEXT{ano}.dbc base_dados_dbc')
```

Figura 3: script para download dos dados

O segundo problema é o formato que o DATASUS disponibiliza os dados, extensão *".dbc"*, um formato proprietário que poucos softwares suportam, em desacordo com o preceituado pelas legislações federais de dados abertos e acesso à informação.

A solução para este problema foi utilizar um pequeno script em R para converter de *".dbc"* para *".csv"*.

```r
library(read.dbc)

# Defina o diretório onde estão os arquivos .dbc
dbc_dir <- "/home/eduardoferreira1626/base_dados_dbc"

# diretorio para salvar os arquivos .csv
csv_dir <- "/home/eduardoferreira1626/base_dados_csv"

# Liste os arquivos .dbc no diretório
dbc_files <- list.files(dbc_dir, pattern = ".dbc", full.names = TRUE)

# Verifique se há arquivos .dbc
if (length(dbc_files) == 0) {
  cat("Nenhum arquivo .dbc encontrado no diretório especificado.\n")
} else {
  # Loop através dos arquivos e converta-os em CSV
  for (file in dbc_files) {
    # Carregue o arquivo .dbc
    dbc <- read.dbc(file)

    # Extraia os dados do arquivo .dbc
    data <- as.data.frame(dbc)

    # Determine o nome do arquivo CSV de saída
    csv_file <- sub(".dbc", ".csv", basename(file))
```

Figura 4: script em R para conversão dbc para csv

Deve-se salientar que para a execução do script houve a necessidade de instalação do R no ambiente da CGP. Quanto a esse ponto, um destaque positivo para a plataforma escolhida foi que o ambiente Linux disponibilizado é baseado em Debian, os procedimentos em linha de comando são comuns para quem tem

familiaridade com Linux, além de possuir uma vasta documentação e tutoriais na internet.

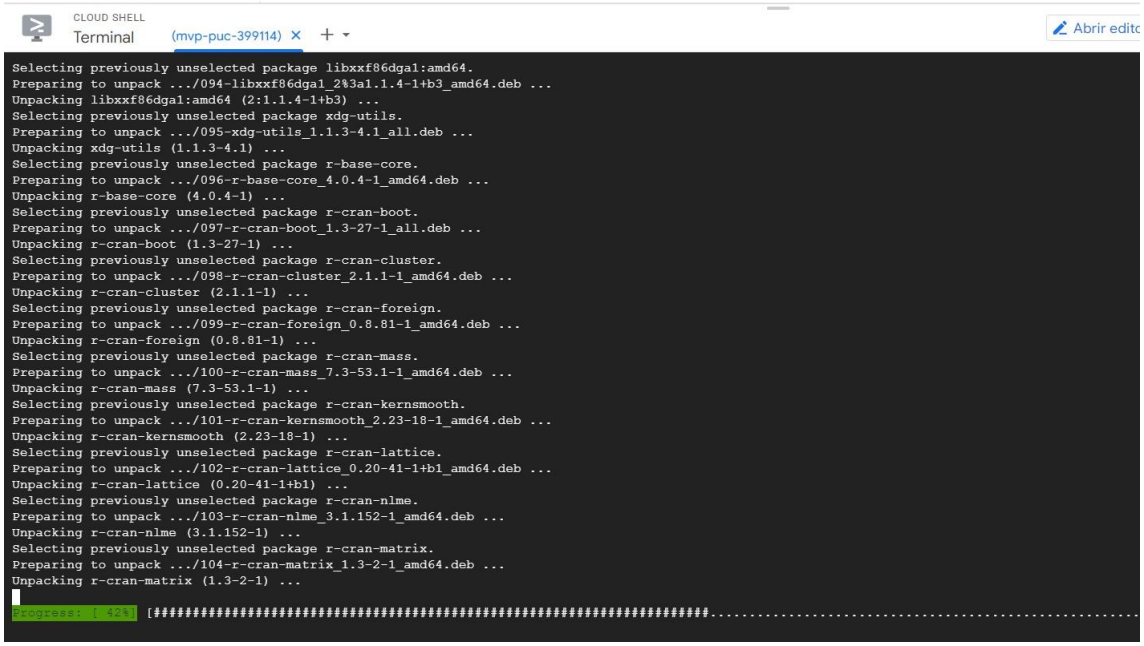


Figura 5: instalação R

VII. Dicionário dos dados

Na página de downloads dos dados, há a opção de se obter documentação relativa aos dados, neste momento obtivemos as legislações sobre o Sistema de Informações de Mortalidade – SIM e o dicionário de dados, que é nomeado pelo DATASUS como “Estrutura do SIM”, com última atualização em dezembro de 2019.

Apresenta-se a seguir o dicionário dos dados com a descrição dos campos, tipos de dados originais, e tamanho das células.

Sessão da DO	Nome do campo	Nome do campo no DBF	Tipo	Tam	Valores válidos	Descrição	Características
I - Identificação	Tipo do óbito	TIPOBITO	Caractere	1	1-Fetal; 2-Não Fetal	Tipo do óbito Óbito fetal: morte antes da expulsão ou da extração completa do corpo da Mãe, independentemente da duração da gravidez. Indica o óbito o fato de o feto, depois da expulsão do corpo materno, não respirar nem apresentar nenhum outro sinal de vida, como batimentos do coração, pulsações do cordão umbilical ou movimentos efetivos dos músculos de contração voluntária	Campo obrigatório
	Data do Óbito	DTOBITO	Caractere	8	Data no padrão ddmmaaaa	Data em que ocorreu o óbito.	Campo obrigatório
	Hora	HORAOBITO	caractere	5	Números (padrão 24 horas 00:00)	Horário do óbito	
	Naturalidade	NATURAL	caractere	3	Números	País e Unidade da Federação onde falecido nasceu. Se estrangeiro informar País	
	Código do município de naturalidade	CODMUNNATU	caractere	7	Números	Código do município de naturalidade do falecido.	
	Data de Nascimento	DTNASC	caractere	8	Data no padrão ddmmaaaa	Data do nascimento do falecido. Em caso de óbito fetal as datas de óbito e nascimento deverão ser iguais.	

	Idade	IDADE	caractere	3	Idade: composto de dois subcampos. - O primeiro, de 1 dígito, indica a unidade da idade (se 1 = minuto, se 2 = hora, se 3 = mês, se 4 = ano, se = 5, idade maior que 100 anos). - O segundo, de dois dígitos, indica a quantidade de unidades: **Idade menor de 1 hora**: subcampo varia de 01 e 59 (minutos); **De 1 a 23 Horas**: subcampo varia de 01 a 23 (horas); **De 24 horas e 29 dias:** subcampo varia de 01 a 29 (dias); **De 1 a menos de 12 meses completos**: subcampo varia de 01 a 11 (meses); **Anos** - subcampo varia de 00 a 99; - 9 - ignorado	Idade do falecido em minutos, horas, dias, meses ou anos	Campo obrigatório. Se 1 = 1 (óbito fetal) campo não deve ser preenchido.
	Sexo	SEXO	caractere	1	M – masculino; F – feminino; I - ignorado	Sexo do falecido. “Ignorado” selecionada em casos especiais como cadáveres mutilados, em estado avançado de decomposição, genitália indefinida ou hermafroditismo	Campo obrigatório
	Raça Cor	RACACOR	caractere	1	1 – Branca; 2 – Preta; 3 – Amarela; 4 – Parda; 5 – Indígena	Cor informada pelo responsável pelas informações do falecido	
	Situação Conjugal	ESTCIV	caractere	1	1 – Solteiro; 2 – Casado; 3 – Viúvo; 4 – Separado judicialmente/divorciado; 5 – União estável; 9 – Ignorado	Situação conjugal do falecido informada pelos familiares	

	Escolaridade (última série concluída) - nível	ESC2010	caractere	1	0 – Sem escolaridade; 1 – Fundamental I (1ª a 4ª série); 2 – Fundamental II (5ª a 8ª série); 3 – Médio (antigo 2º Grau); 4 – Superior incompleto; 5 – Superior completo; 9 – Ignorado	Escolaridade 2010. Nível da última série concluída pelo falecido	
	Escolaridade (última série concluída) - série	SERIESCFAL	caractere		Números de 1 a 8	Última série escolar concluída pelo falecido	
	Ocupação habitual (Código CBO 2002)	OCUP	caractere	6	Números	Tipo de trabalho que o falecido desenvolveu na maior parte de sua vida produtiva. Preenchimento de acordo com Classificação Brasileira de Ocupações – CBO 2002	Campo não preenchido para óbitos em crianças com menos de 5 anos
II - Residência	Município de residência (Código)	CODMUNRES	caractere	7	Números	Código do município de residência. Em caso de óbito fetal, considerar o município de residência da mãe	Campo obrigatório
III- Ocorrência	Local de Ocorrência do Óbito	LOCOCOR	caractere	1	1 – hospital; 2 – outros estabelecimentos de saúde; 3 – domicílio; 4 – via pública; 5 – outros; 6 - aldeia indígena; 9 – ignorado.	Local de ocorrência do óbito	Campo obrigatório
	Código do Município de ocorrência	CODMUNOCOR	caractere	8	Números	Código relativo ao município onde ocorreu o óbito	Campo obrigatório

IV – Fetal ou menor que 1 ano	Idade (anos)	IDADEMAE	caractere	2	Números	Idade da mãe	
	Escolaridade (última série concluída) - nível	ESCMAE2010	caractere	1	0 – Sem escolaridade; 1 – Fundamental I (1ª a 4ª série); 2 – Fundamental II (5ª a 8ª série); 3 – Médio (antigo 2º Grau); 4 – Superior incompleto; 5 – Superior completo; 9 – Ignorado	Escolaridade 2010. Nível da última série concluída pela mãe	
	Escolaridade (última série concluída) - série	SERIESCMAE	caractere	1	Números de 1 a 8	Última série escolar concluída pela mãe	Campo preenchido se 28 = 1, 2 ou 3
	Ocupação habitual (Código CBO 2002)	OCUPMAE	caractere	6		Tipo de trabalho exercido habitualmente pela Mãe, de acordo com Classificação Brasileira de Ocupações – CBO 2002. No caso da mãe do falecido(a) ser "aposentada", preencher com a ocupação habitual anterior.	
	Número de filhos tidos (nascidos vivos)	QTDFILVIVO	caractere	2	Número; 9 - ignorado	Número de filhos vivos	
	Número de filhos tidos (perdas fetais/aborto)	QTDFILMORT	caractere	2	Número; 9 - ignorado	Número de filhos mortos. Não incluir a criança cujo óbito se notifica na respectiva DO	
	Nº de semanas de gestação	SEMAGESTAC	caractere	3	Números com dois algarismos; 9 - ignorado	Semanas de gestação com dois algarismos	
	Tipo de Gravidez	GRAVIDEZ	caractere	1	1 – única; 2 – dupla; 3 – tripla e mais; 9 – ignorada	Tipo de gravidez	
	Tipo de Parto	PARTO	caractere	1	1 – vaginal; 2 – cesáreo; 9 – ignorado	Tipo de parto	

	Morte em relação ao Parto	OBITOPARTO	caractere	1	1 - antes; 2– durante; 3–depois; 9– Ignorado	Momento do óbito em relação ao parto	
	Peso ao Nascer	PESO	caractere	4	Número (quatro algarismos)	Peso ao nascer em gramas	
	Número da Declaração de Nascido Vivo	NUMERODN	caractere	8	Números	Número da Declaração de Nascido Vivo	
V – Condições e causas do óbito	A morte ocorreu	TPMORTEOCO	caractere	1	1 – na gravidez; 2 – no parto; 3 – no abortamento; 4 – até 42 dias após o término do parto; 5 – de 43 dias a 1 ano após o término da gestação; 8 – não ocorreu nestes períodos; 9 – ignorado.	Situação gestacional ou pós gestacional em que ocorreu o óbito	Deve ser preenchido em caso de óbito de mulher fértil
	Recebeu assist. médica durante a doença que ocasionou a morte?	ASSISTMED	caractere	1	1 – sim; 2 – não; 9 – ignorado	Se refere ao atendimento médico continuado que o paciente recebeu, ou não, durante a enfermidade que ocasionou o óbito	
	Necrópsia	NECROPSIA	caractere	1	1 – sim; 2 – não; 9 – ignorado	Refere-se a execução ou não de necropsia para confirmação do diagnóstico	
	Causas da Morte- Parte I - CID	LINHAA	caractere	20	Códigos CID 10	CIDs informados na Linha A da DO referente ao diagnóstico na Linha A da DO (**causa terminal -** doença ou estado mórbido que causou diretamente a morte)	Campo obrigatório
	Causas da Morte- Parte I - CID	LINHAB	caractere	20	Códigos CID 10	CIDs informados na Linha B da DO referente ao diagnóstico na Linha B da DO (**causa antecedente ou consequencial** - estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A)	

	Causas da Morte- Parte I - CID	LINHAC	caractere	20	Códigos CID 10	CIDs informados na Linha C da DO referente ao diagnóstico na Linha C da DO (**causa antecedente ou consequencial** - estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A)	
	Causas da Morte- Parte I - CID	LINHAD	caractere	20	Códigos CID 10	CIDs informados na Linha D da DO referente ao diagnóstico na Linha D da DO **(causa básica** – estado mórbido, se existir, que produziu a causa direta da morte registrada na linha A)	
	Causas da Morte- Parte II - CID	LINHAII	caractere	45	Códigos CID 10	CIDs informados na Parte II da DO referente ao diagnóstico na Parte II da DO (**causa contribuinte** - outras condições significativas que contribuíram para a morte e que não entraram na cadeia definida na Parte I	
VII – Causas externas	Tipo	CIRCOBITO	caractere	1	1 – acidente; 2 – suicídio; 3 – homicídio; 4 – outros; 9 – ignorado	Tipo de morte violenta ou circunstâncias em que se deu a morte não natural	
	Acidente do trabalho	ACIDTRAB	caractere	1	1 – sim; 2 – não; 9 – ignorado	Indica se o evento que desencadeou o óbito está relacionado ao processo de trabalho	
	Fonte da Informação	FONTE	caractere	1	1 – ocorrência policial; 2 – hospital; 3 – família; 4 – outra; 9 – ignorado	fonte de informação utilizada para o preenchimento dos campos 48 e 49	
Demais campos		ORIGEM	caractere	1	1- Oracle; 2 - Banco estadual disponibilizado via FTP; 3 - Banco SEADE; 9 - Ignorado	Origem do registro	Campo calculado pelo sistema
		ESC	caractere	1	1 – Nenhuma; 2 – de 1 a 3 anos; 3 – de 4 a 7 anos; 4 – de 8 a 11 anos; 5 – 12 anos e mais; 9 – Ignorado.	Escolaridade em anos	DO antiga
		ESCMAE	caractere	1	1 – Nenhuma; 2 – de 1 a 3 anos; 3 – de 4 a 7 anos; 4 – de 8 a 11 anos; 5 – 12 anos e mais; 9 – Ignorado	Escolaridade da mãe em anos	DO antiga

	OBITOGRAV	caractere	1	1 – sim; 2 – não; 9 – ignorado	Óbito na gravidez	Ficha de investigação de óbito de mulher em idade fértil
	OBITOPUERP	caractere	1	1 – Sim, até 42 dias após o parto; 2 – Sim, de 43 dias a 1 ano; 3 – Não; 9 – Ignorado.	Óbito no puerpério	Ficha de investigação de óbito de mulher em idade fértil
	EXAME	caractere	1	1 – sim; 2 – não; 9 – ignorado	Realização de exame	Ficha de investigação de óbito materno ou infantil
	CIRURGIA	caractere	1	1 – sim; 2 – não; 9 – ignorado	Realização de cirurgia	Ficha de investigação de óbito materno ou infantil
	CAUSABAS_O	caractere	4	Códigos CID 10	Causa básica informada antes da resseleção	
	CAUSABAS	caractere	4	Códigos CID 10	Causa básica da DO	
	NUMEROLOTE	caractere	8	Números	Número do lote	Campo calculado pelo sistema
	DTINVESTIG	caractere	8	Data no padrão ddmmaaaa	Data da investigação do óbito	Ficha de investigação de óbito materno ou infantil
	DTCADASTRO	caractere	8	Data no padrão ddmmaaaa	Data do cadastro do óbito	Ficha de investigação de óbito materno ou infantil
	STCODIFICA	caractere	1	Se codificadora (valor: S) ou não (valor: N)	Status de instalação	Campo calculado pelo sistema
	CODIFICADO	caractere	1	Se estiver codificado (valor: S) ou não (valor: N)	Informa se formulário foi codificado	Campo calculado pelo sistema
	VERSAOSIST	caractere	7	Números;	Versão do sistema	Campo calculado pelo sistema
	VERSAOSCB	caractere	7	Números;	Versão do seletor de causa básica	Campo calculado pelo sistema

	FONTEINV	caractere	8	1 – Comitê de Morte Materna e/ou Infantil; 2 – Visita domiciliar / Entrevista família; 3 – Estabelecimento de Saúde / Prontuário; 4 – Relacionado com outros bancos de dados; 5 – S V O; 6 – I M L; 7 – Outra fonte; 8 – Múltiplas fontes; 9 – Ignorado	Fonte de investigação	Ficha de investigação de óbito materno ou infantil
	DTRECEBIM	caractere	8	Data no padrão ddmmaaaa	Data do recebimento	Ficha de investigação de óbito materno ou infantil
	ATESTADO	caractere	70	Códigos CID 10	CIDs informados no atestado	
	DTRECORIGA	caractere	8	Data no padrão ddmmaaaa	Campo Criado no Tratamento para Data do recebimento original	Campo calculado pelo sistema
	CAUSAMAT	caractere	4	Códigos CID 10	CID da causa externa associada a uma causa materna	Campo calculado pelo sistema
	ESCMAEAGR1	caractere	2	00 – Sem Escolaridade; 01 – Fundamental I Incompleto; 02 – Fundamental I Completo; 03 – Fundamental II Incompleto; 04 – Fundamental II Completo; 05 – Ensino Médio Incompleto; 06 – Ensino Médio Completo; 07 – Superior Incompleto; 08 – Superior Completo; 09 – Ignorado; 10 – Fundamental I Incompleto ou Inespecífico; 11 – Fundamental II Incompleto ou Inespecífico; 12 – Ensino Médio Incompleto ou	Escolaridade da mãe agregada (formulário a partir de 2010)	

				Inespecífico		
	ESCFALAGR1	caractere	1	00 – Sem Escolaridade; 01 – Fundamental I Incompleto; 02 – Fundamental I Completo; 03 – Fundamental II Incompleto; 04 – Fundamental II Completo; 05 – Ensino Médio Incompleto; 06 – Ensino Médio Completo; 07 – Superior Incompleto; 08 – Superior Completo; 09 – Ignorado; 10 – Fundamental I Incompleto ou Inespecífico; 11 – Fundamental II Incompleto ou Inespecífico; 12 – Ensino Médio Incompleto ou Inespecífico	Escolaridade do falecido agregada (formulário a partir de 2010)	
	STDOEPIDEM	caractere	1	1 - Sim; 0 - Não	Status de DO Epidemiológica	Campo calculado pelo sistema
	STDONOVA	caractere	1	1 - Sim; 0 - Não	Status de DO Nova	Campo calculado pelo sistema

	DIFDATA	caractere	8	Números	Diferença entre a data de óbito e data do recebimento original da DO ([DTOBITO] – [DTRECORIG])	Campo calculado pelo sistema
	NUDIASOBCO	caractere	4	Números	Diferença entre a data óbito e a data conclusão da investigação, em dias.	Campo calculado pelo sistema
	DTCADINV	caractere	8	Data no padrão ddmmaaaa	Data do cadastro de investigação	Ficha de investigação de óbito materno ou infantil
	TPOBITOCOR	caractere	1	1-Durante a gestação, 2- Durante o abortamento, 3- Após o abortamento, 4- No parto ou até 1 hora após o parto, 5- No puerpério - até 42 dias após o parto, 6- Entre 43 dias e até 1 ano após o parto, 7- A investigação não identificou o momento do óbito, 8- Mais de um ano após o parto, 9- O óbito não ocorreu nas circunstâncias anteriores, Branco - Não investigado	Momento da ocorrência do óbito	Ficha de investigação de óbito materno
	DTCONINV	caractere	8	Data no padrão ddmmaaaa	Data da conclusão da investigação	Ficha de investigação de óbito materno ou infantil
	FONTES	caractere	6	Letras	Combinado de caracteres conforme o preenchimento dos campos de fontes (FONTENTREV, FONTEAMBUL, FONTEPRONT, FONTESVO, FONTEIML, FONTEPROF): se preenchido caractere “S”, se o campo estiver vazio caractere “X”	Campo calculado pelo sistema

		TPRESGINFO	caractere	2	01 - Não acrescentou nem corrigiu informação; 02 - Sim, permitiu o resgate de novas informações; 03 - Sim, permitiu a correção de alguma das causas informadas originalmente.	Informa se a investigação permitiu o resgate de alguma causa de óbito não informado, ou a correção de alguma antes informada	Campo calculado pelo sistema
		TPNIVELINV	caractere	1	E – estadual; R- regional; M- Municipal	Tipo de nível investigador	Campo calculado pelo sistema
		DTCADINF	caractere	8	Data no padrão ddmmaaaa	Quando preenchido indica se a investigação foi realizada	Ficha de investigação de óbito materno ou infantil
		MORTEPARTO	caractere	1	1 - antes; 2– durante; 3–após; 9– Ignorado	Momento do óbito em relação ao parto	
		DTCONCASO	caractere	8	Data no padrão ddmmaaaa	Data de conclusão do caso	Ficha de investigação de óbito materno ou infantil
		ALTCAUSA	caractere	1	1- Sim; 2 – Não	Indica se houve correção ou alteração da causa do óbito após investigação	Ficha de investigação de óbito materno ou infantil
		TPPOS	caractere	1	1 – sim; 2 – não	Óbito investigado	Campo calculado pelo sistema
	Semanas de gestação (formulário antigo)	GESTACAO	caractere	1	1 - Menos de 22 semanas; 2 - 22 a 27 semanas; 3 - 28 a 31 semanas; 4 - 32 a 36 semanas; 5 - 37 a 41 semanas; 6 - 42 e + semanas	Faixas de semanas de gestação	
		CB_PRE	caractere	4	Código CID 10	Causa básica informada antes da resseleção (localidade)	Ficha de investigação de óbito materno ou infantil

VIII. Tabelas auxiliares

Conforme pode se observar no dicionário dos dados, alguns dados ficam sem significado, obrigando a se obter a informação faltante em outras fontes.

A informação sobre a ocupação, a Classificação Brasileira de Ocupações CBO, foi obtida no site do Ministério do Trabalho.

O download foi manual porque o site tem proteção de download automático.

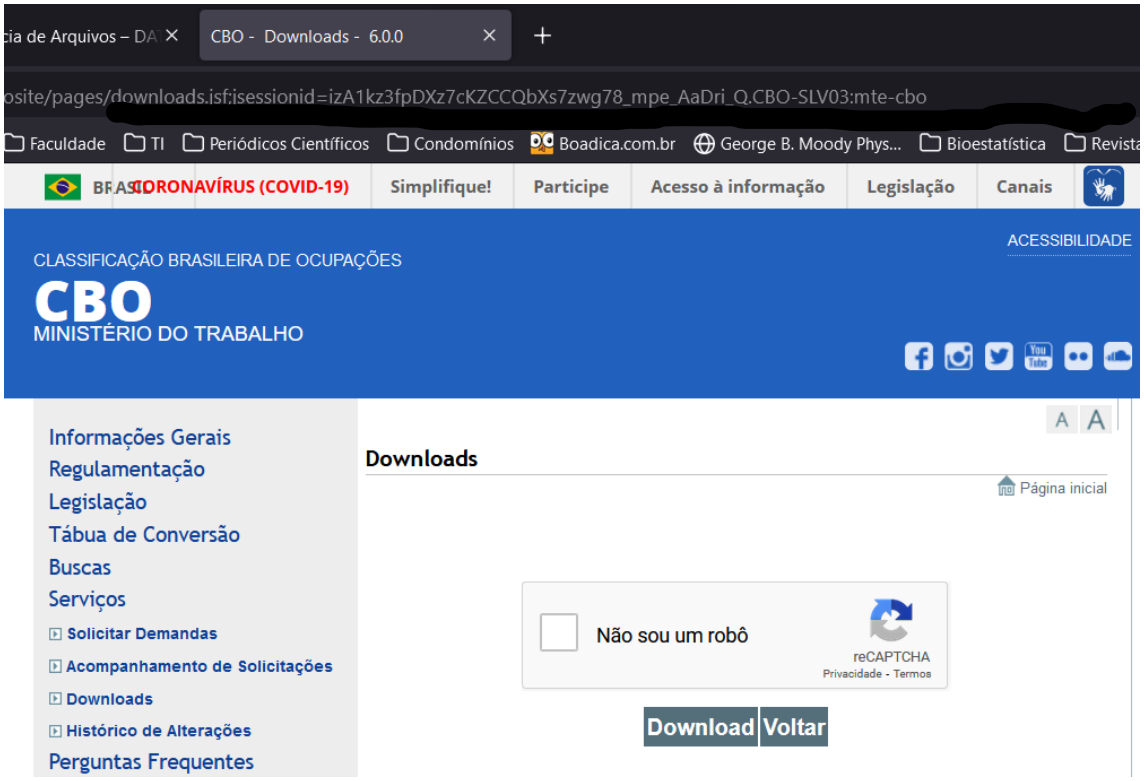


Figura 6: Site Ministério do Trabalho

Dos arquivos disponibilizados no arquivo compactado, utilizou-se o arquivo CBO2002 – Ocupacao.csv que contém 2 colunas:

Nome do campo	Tipo	Tamanho
CODIGO	Inteiro	5 a 6 dígitos
TITULO	Caractere	4 a 136 letras

Destacando que para a utilização não tivesse problemas, foi necessária uma transformação de codificação do arquivo csv para o padrão utilizado no GCP.

```
import pandas as pd

dados = pd.read_csv('/home/eduardoferreira1626/temporario/CBO2002 - Ocupacao.csv', encoding='ISO-8859-1', sep=';')

#dados = dados.drop_duplicates()

dados.to_csv('dados_auxiliares/tabela_ocupacoes.csv', index=False)
```

Figura 7: script mudança de codificação csv

A informação dos códigos de município e unidades da federação foi alcançada através da tabela disponibilizada pelo IBGE, obtida por download pelo comando *wget*, e descompactação pelo comando *unzip* direto no Cloud Shell.

```
eduardoferreira1626@cloudshell:~ (mvp-puc-399114)$ wget https://geoftp.ibge.gov.br/organizacao_do_territorio/estrutura_territ
orial/divisao_territorial/2022/DTB_2022.zip
--2023-09-25 15:57:52--  https://geoftp.ibge.gov.br/organizacao_do_territorio/estrutura_territorial/divisao_territorial/2022/
DTB_2022.zip
Resolving geoftp.ibge.gov.br (geoftp.ibge.gov.br)... 170.84.40.11
Connecting to geoftp.ibge.gov.br (geoftp.ibge.gov.br)|170.84.40.11|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2039162 (1.9M) [application/zip]
Saving to: 'DTB_2022.zip'

DTB_2022.zip                    100%[===================================================>]   1.94M   792KB/s    in 2.5s

2023-09-25 15:57:56 (792 KB/s) - 'DTB_2022.zip' saved [2039162/2039162]

eduardoferreira1626@cloudshell:~ (mvp-puc-399114)$ unzip DTB_2022.zip
Archive:  DTB_2022.zip
  inflating: RELATORIO_DTB_BRASIL_DISTRITO.ods
  inflating: RELATORIO_DTB_BRASIL_DISTRITO.xls
  inflating: RELATORIO_DTB_BRASIL_MUNICIPIO.ods
  inflating: RELATORIO_DTB_BRASIL_MUNICIPIO.xls
  inflating: RELATORIO_DTB_BRASIL_SUBDISTRITO.ods
  inflating: RELATORIO_DTB_BRASIL_SUBDISTRITO.xls
eduardoferreira1626@cloudshell:~ (mvp-puc-399114)$
```

Figura 8: download arquivos IBGE

Como a informação da Unidade da Federação – UF é suficiente para o objetivo deste trabalho e nos dados disponibilizados na tabela do DATASUS a informação do local onde ocorreu o óbito CODMUNOCOR é uma sequência numérica com 8 dígitos e nas tabelas do IBGE os municípios são códigos numéricos de 5 ou 7 dígitos, foi feita uma adequação do arquivo RELATORIO_DTB_BRASIL_MUNICIPIO.xls para que tivesse apenas o código da UF e o nome, além de converter em um arquivo csv.

```
import pandas as pd

dados = pd.read_excel('/home/eduardoferreira1626/temporario/RELATORIO_DTB_BRASIL_MUNICIPIO.xls', skiprows=6)

dados = dados[['UF', 'Nome_UF']]

dados = dados.drop_duplicates()

dados.to_csv('dados_auxiliares/ibge_codigo_UF.csv', index=False)
```

Figura 9: script modificação dados IBGE

A terceira tabela com informações foi a descrição da causa morte, conforme a documentação dos dados, o campo CAUSABAS é o código da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde - CID-10. O arquivo csv é disponibilizado pelo próprio DATASUS na parte que trata do CID-10.

O download destes arquivos também foi manual porque a conexão expirava sem terminar o download pelo link disponibilizado, mas pelo navegador foi sem problemas.

A dinâmica de obtenção foi o download do arquivo compactado com posterior descompactação das tabelas e envio para o ambiente do GCP.

```
CLOUD SHELL
Terminal (mvp-puc-399114)                                   Abrir editor
ders.
Retrying.

--2023-09-25 17:14:07--  (try: 3)  http://www2.datasus.gov.br/cid10/V2008/downloads/CID10CSV.zip
Connecting to www2.datasus.gov.br (www2.datasus.gov.br)|189.28.139.128|:80... connected.
HTTP request sent, awaiting response... Read error (Connection timed out) in headers.
Retrying.

--2023-09-25 17:29:11--  (try: 4)  http://www2.datasus.gov.br/cid10/V2008/downloads/CID10CSV.zip
Connecting to www2.datasus.gov.br (www2.datasus.gov.br)|189.28.139.128|:80... connected.
HTTP request sent, awaiting response... Read error (Connection timed out) in headers.
Retrying.

--2023-09-25 17:44:15--  (try: 5)  http://www2.datasus.gov.br/cid10/V2008/downloads/CID10CSV.zip
Connecting to www2.datasus.gov.br (www2.datasus.gov.br)|189.28.139.128|:80... connected.
HTTP request sent, awaiting response...
```

Figura 10: tentativa download dados CID-10

Quanto a adequação da tabela para a utilização (filtragem das colunas necessárias e codificação csv), foi utilizado mais um script em Python.

```
CLOUD SHELL
Editor                                                      Abrir terminal
File  Edit  Selection  View  Go  Run  Terminal  Help
EXPLORER                    conversao_tabela_cid10.py
> OPEN EDITORS              scripts > conversao_tabela_cid10.py > ...
EDUARDOFERREIRA1626
> base_dados_csv            import pandas as pd
> base_dados_dbc
> base_dados_pronta         dados = pd.read_csv('/home/eduardoferreira1626/temporario/CID-10-SUBCATEGORIAS.CSV', encoding='ISO-8859-1', sep=';')
> dados_auxiliares
  ibge_codigo_UF.csv        dados = dados[['SUBCAT', 'DESCRICAO']]
  tabela_cid10.csv
                            dados.to_csv('/home/eduardoferreira1626/dados_auxiliares/tabela_cid10.csv', index=False)
```

Figura 11: script modificação e codificação csv CID-10

O que deixou a tabela s ser utilizada da seguinte maneira:

Nome do campo	Tipo	Tamanho
SUBCAT	Caractere	3 a 4 letras
DESCRICAO	Caractere	5 a 264 letras

IX. Modelagem

Considerando a pretensão inicial do estudo de gerar um perfil das pessoas que cometem suicídio, entendeu-se que a arquitetura mais conveniente para o consumo dos dados seria a Flat Table, ou seja, uma tabela desnormalizada. No final do ciclo ETL, teremos uma tabela única com todas as informações.

X. Armazenamento

A infraestrutura da GCP para o armazenamento dos dados é a Google Cloud Storage, através de Buckets.

Foram criados 2 buckets para o armazenamento dos dados obtidos e um foi criado automaticamente pela ferramenta de ETL DataPrep, que será tratada posteriormente.

O bucket chamado de “dados-complementares” ficou destinado às tabelas auxiliares obtidos no Ministério do Trabalho, IBGE e no próprio DATASUS que seriam usadas para complementar as informações da tabela principal.

O espaço principal foi pensado em armazenar a série histórica das informações de óbitos de causas externas, chamado de “óbitos-datasus”.

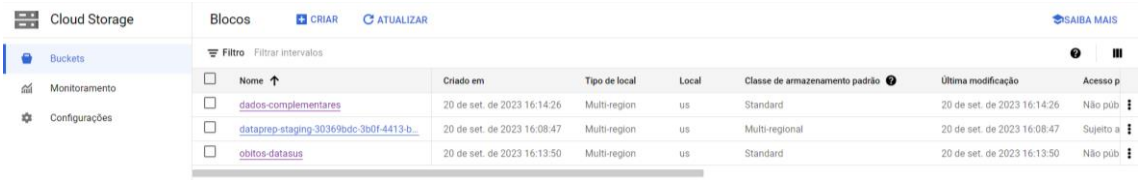


Figura 12: Lista buckets

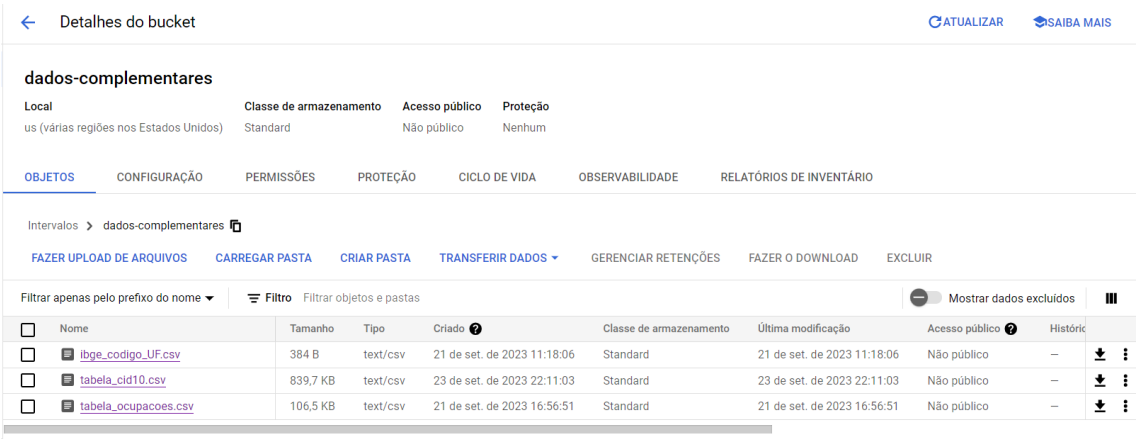


Figura 13: bucket dados-complementares

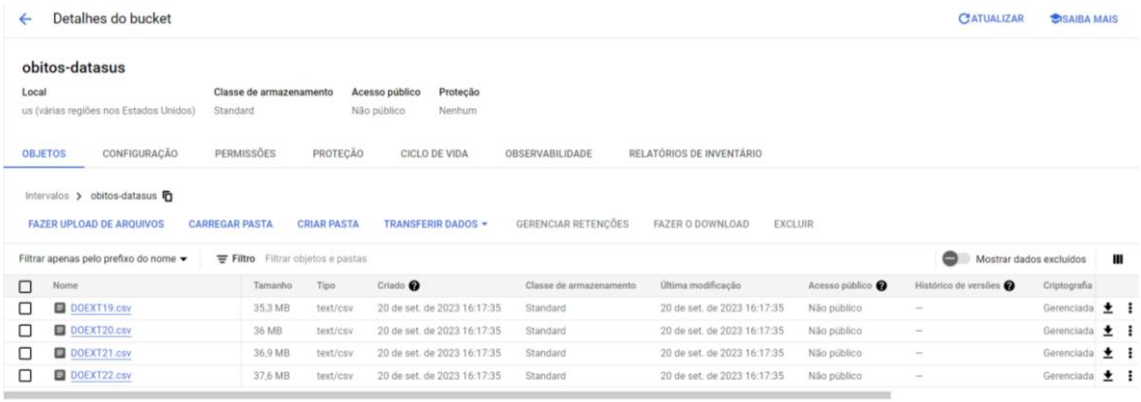


Figura 14: bucket obitos-datasus

Cabe pontuar que a transferência dos dados com o tratamento inicial para os buckets foi feita através de linha de comando no ambiente Cloud Shell e enviados para os buckets correspondentes.

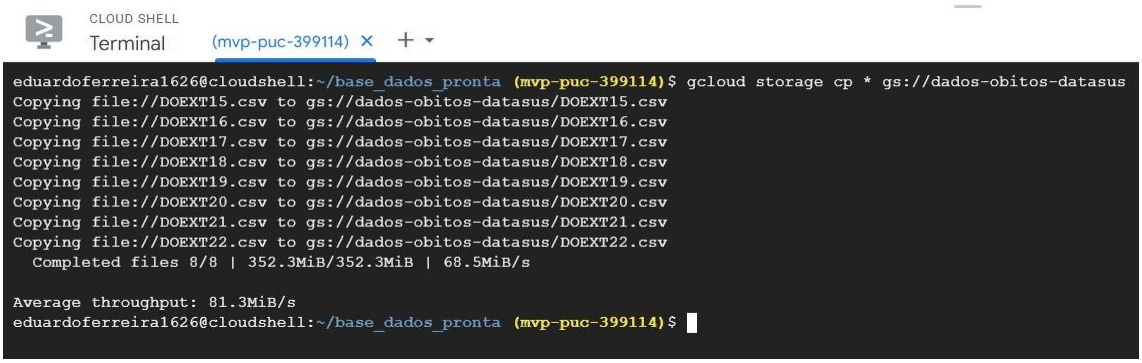


Figura 15: transferência dados para bucket

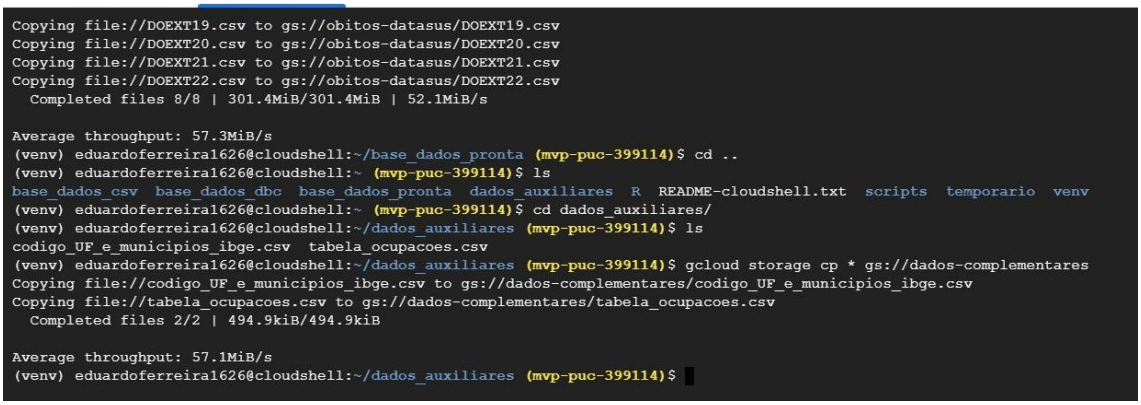


Figura 16: transferência dados bucket dados auxiliares

XI. Transformação

Para o fluxo de transformação ETL foi escolhido inicialmente a ferramenta Cloud Data Fusion, porém depois de muitas tentativas sem sucesso.

Embora a ferramenta não informasse inconsistências nas configurações dos modelos de modificação do Wrangler, não houve sucesso em manipular os dados pelo Data Fusion.

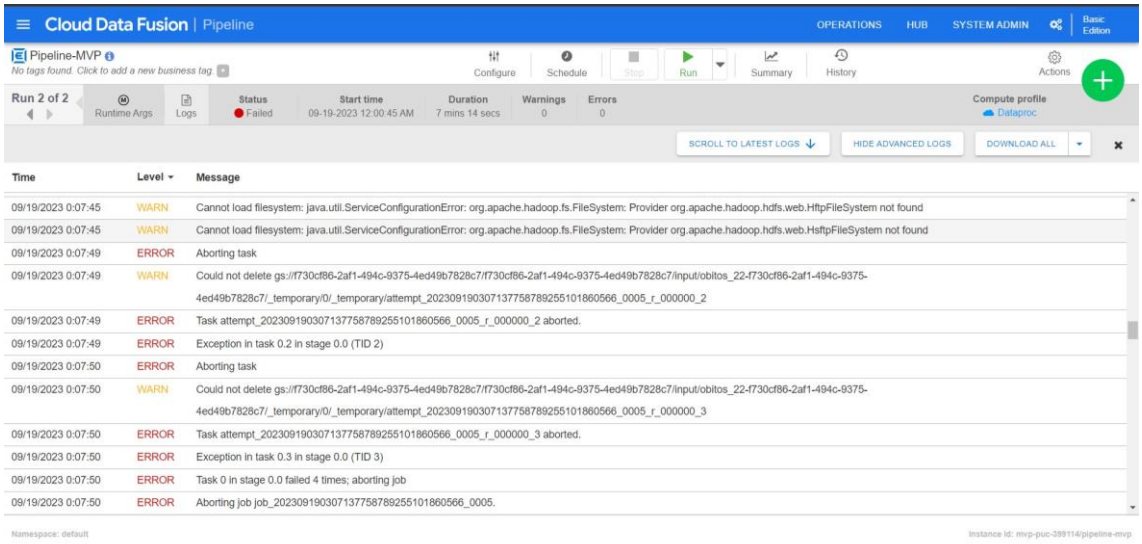


Figura 17: log de erro Data Fusion

Destaca-se que as tentativas no Data Fusion foram dos fluxos mais complexos até o mais simples, conforme figuras a seguir.

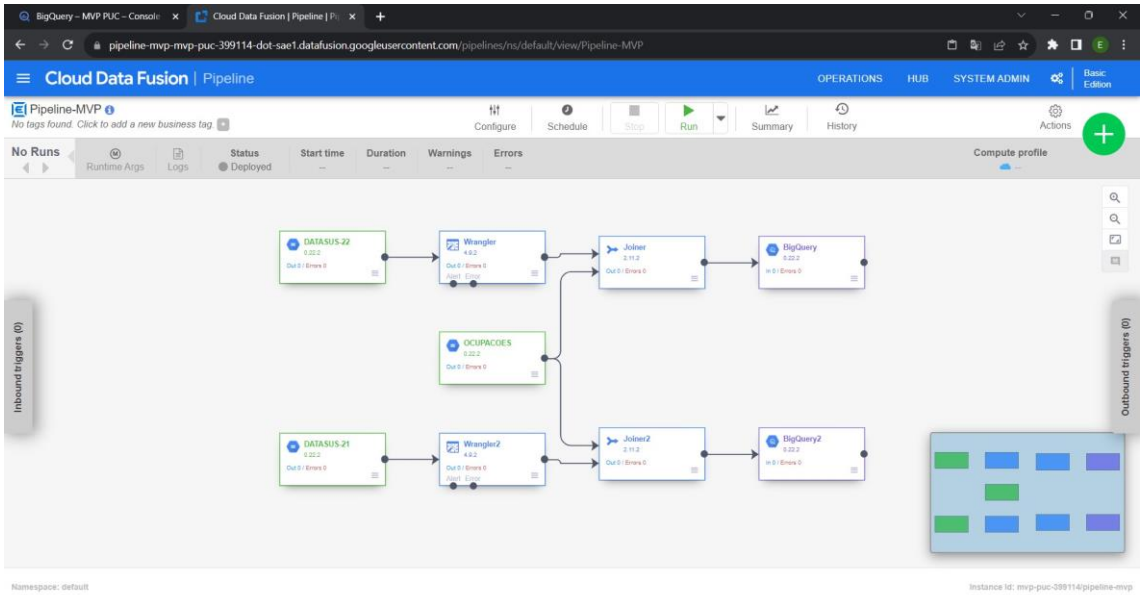


Figura 18: fluxo inicial Data Fusion

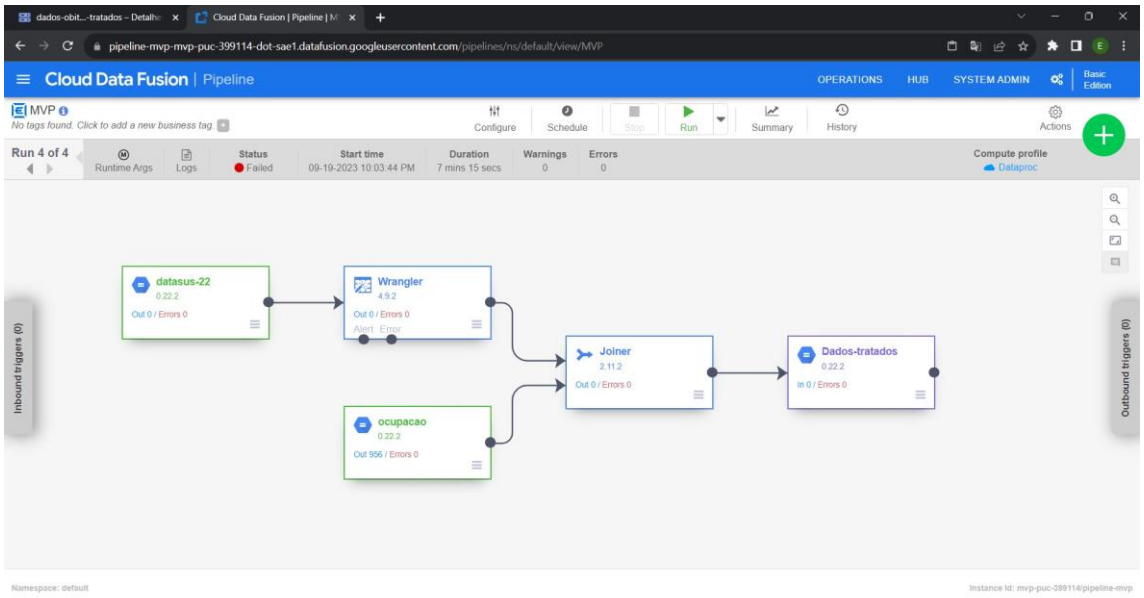


Figura 19: fluxo intermediário Data Fusion

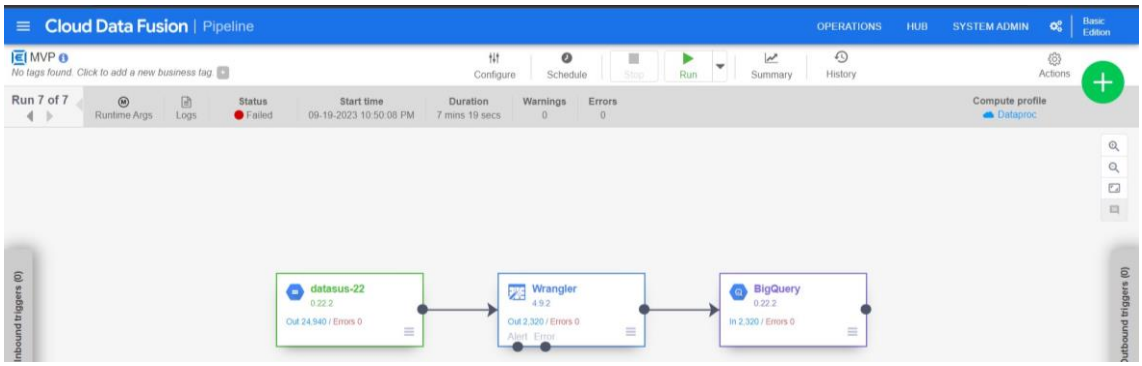


Figura 20: fluxo simplificado Data Fusion

Com o insucesso de implementar um fluxo de transformação através do Data Fusion, encontramos outra ferramenta com a proposta similar: o Dataprep, que é uma ferramenta com interface *low code* que implementa e agenda fluxos de tratamento de dados com a infraestrutura do Dataflow.

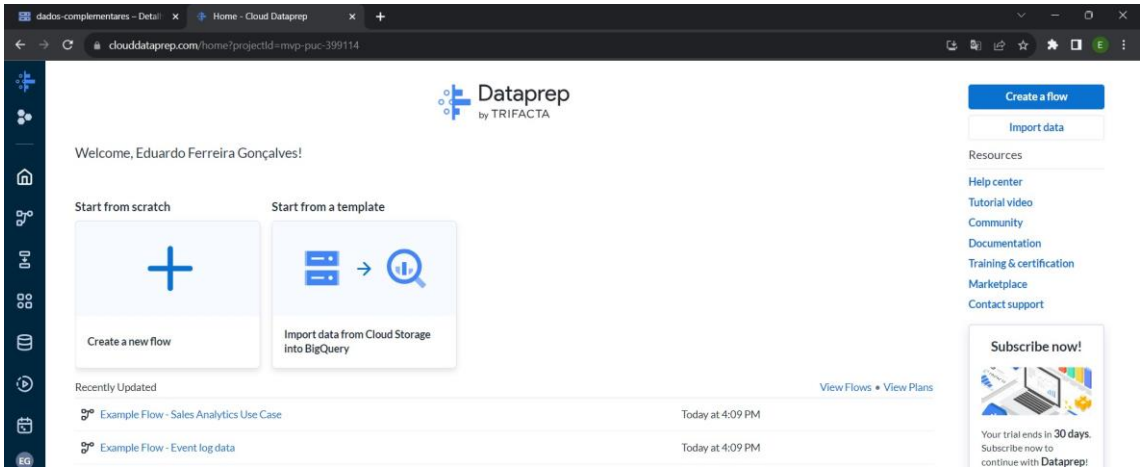


Figura 21: Tela inicial Dataprep

O fluxo proposto neste momento foi a obtenção dos dados no bucket da Cloud Storage, manipulação conforme receita gerada e finalização do dado tratado com a disponibilização no BigQuery.

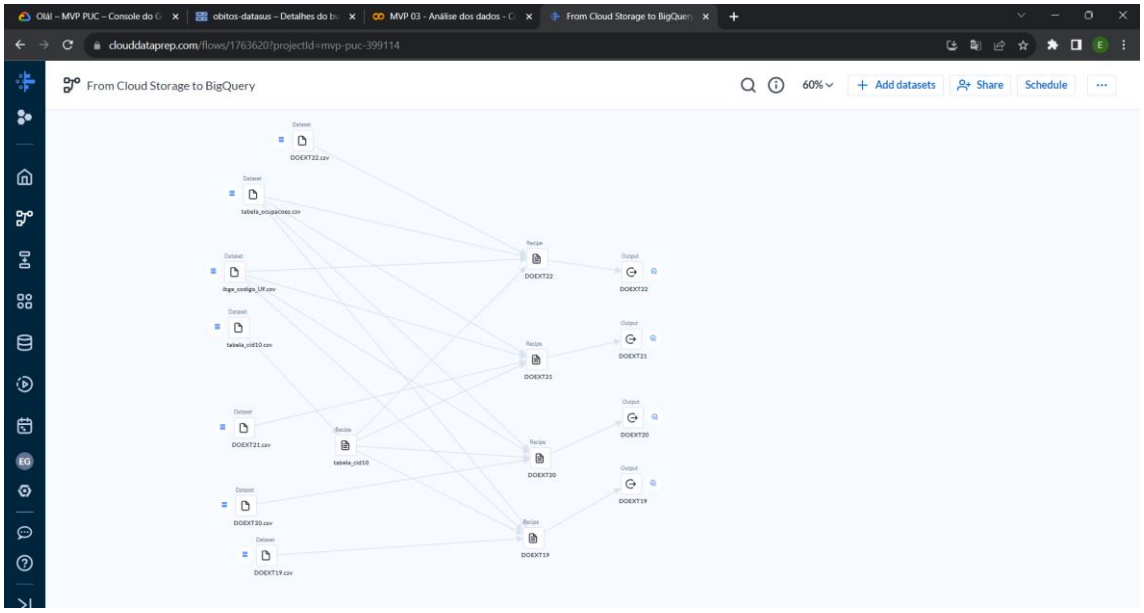


Figura 22: Fluxo dados Dataprep

Para cada fonte ou conjunto de fonte de dados que serão manipulados para a posterior disponibilização cria um ponto de saída, com o sei passo de receita particular.

O fluxo demonstrado na figura foi desenhado pela entrada de 4 anos de dados dos óbitos por causas externas (DOEXT) e as tabelas auxiliares de Ocupações, Unidades federativas e CID-10.

Os nós *recipe* (DOEXT19, DOEXT20, DOEXT21 e DOEXT22) se responsabilizam pelo tratamento dos dados das fontes e das junções das tabelas.

Com uma interface intuitiva similar ao Microsoft Excel, a ferramenta obtém o schema de forma automática, permite a modificação dos tipos de dados de cada variável, além disponibilizar formas de tratamento de valores faltantes etc.

Destacando que para cada manipulação nos dados é adicionado um step na lista da receita de manipulação. Como os dados têm as mesmas variáveis/colunas, pode-se aproveitar a receita para todos os anos.

Como a ferramenta possibilita o download da íntegra da receita de manipulação feita, disponibilizamos ao final deste trabalho a íntegra da manipulação executada através da ferramenta.

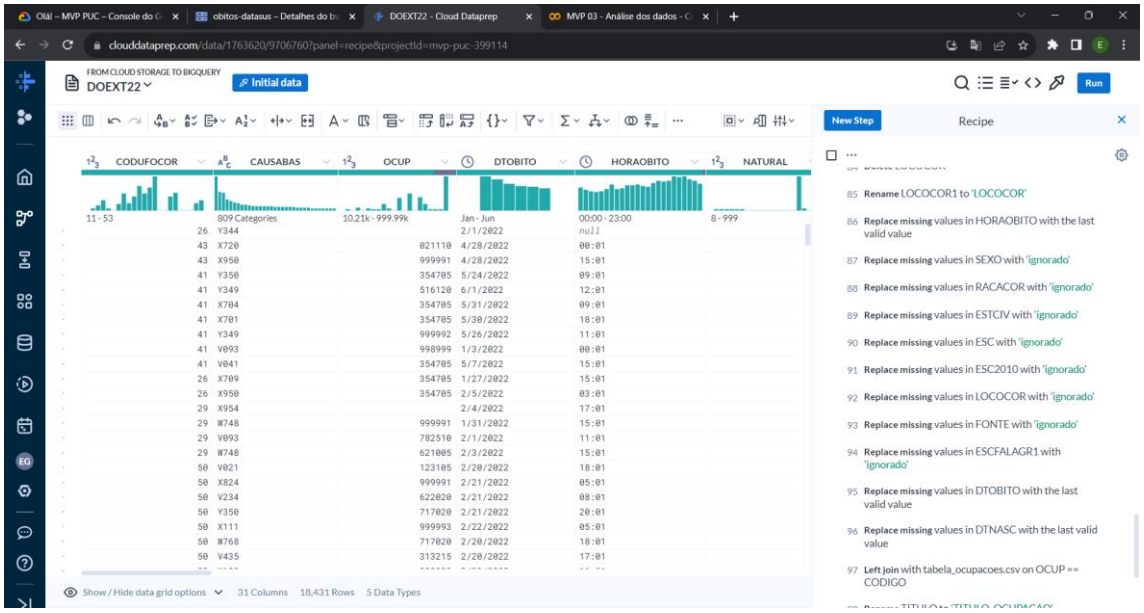


Figura 23: tela do wrangler Dataprep

Após os fluxos preparados e salvos, o Dataprep cria um job no Dataflow e executa as transformações de forma agendada ou sob demanda.

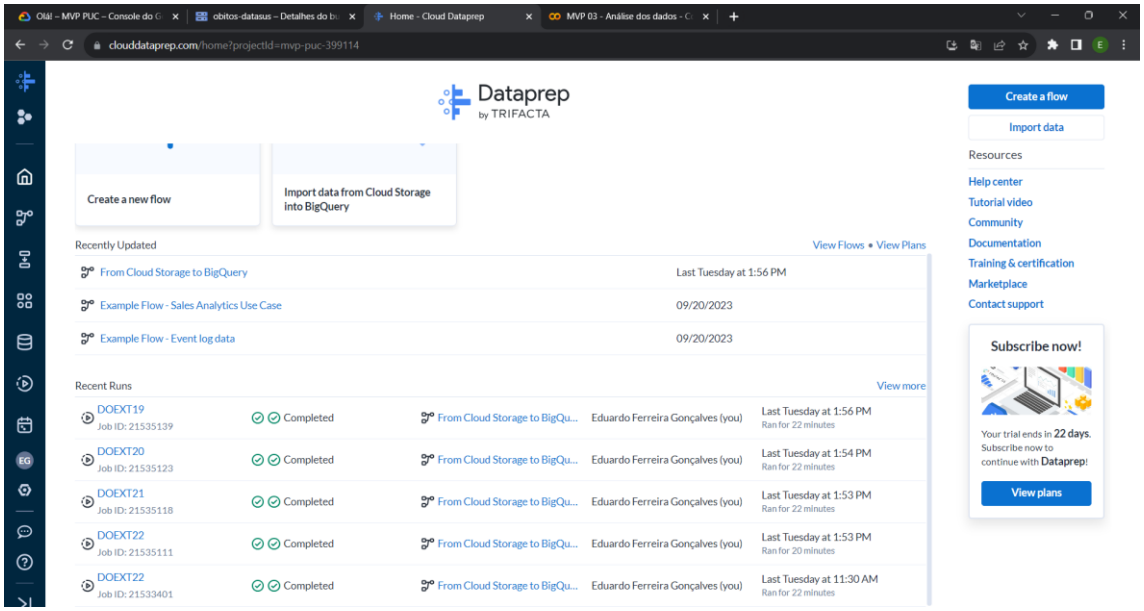


Figura 24: lista das execuções realizadas

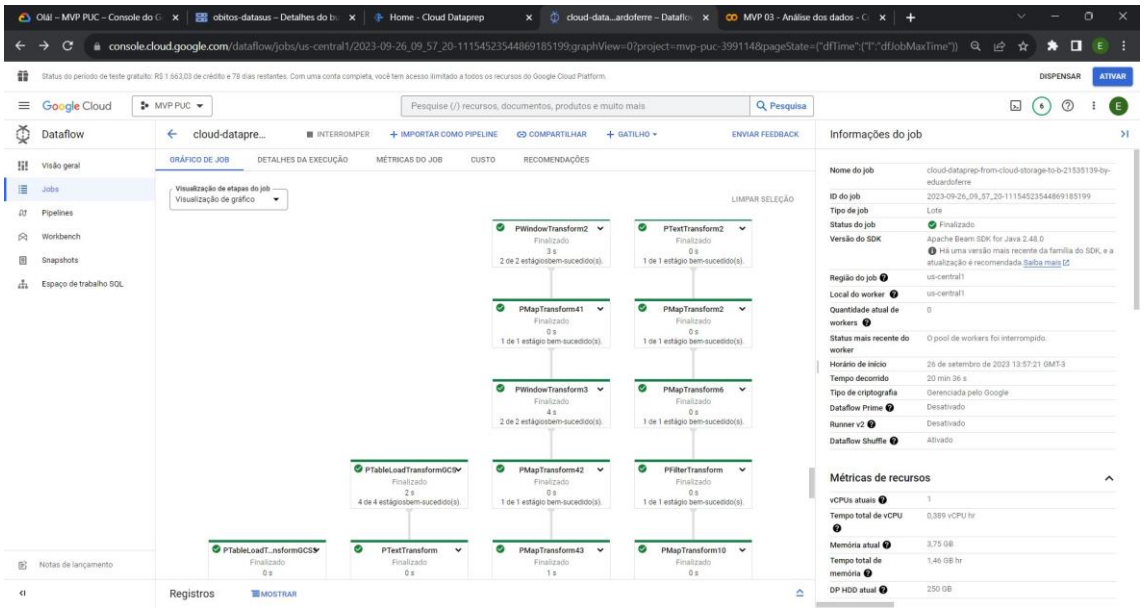


Figura 25: Fluxo de transformação Dataflow

XII. Disponibilização

Ao finalizar o processamento do pipeline de extração, transformação, é o momento de carregar os dados. A ferramenta da GCP para este fim é o BigQuery.

Destacando que antes do carregamento por parte do Dataprep, criou-se um dataset no BigQuery para que os dados fossem carregados como tabelas dentro do dataset.

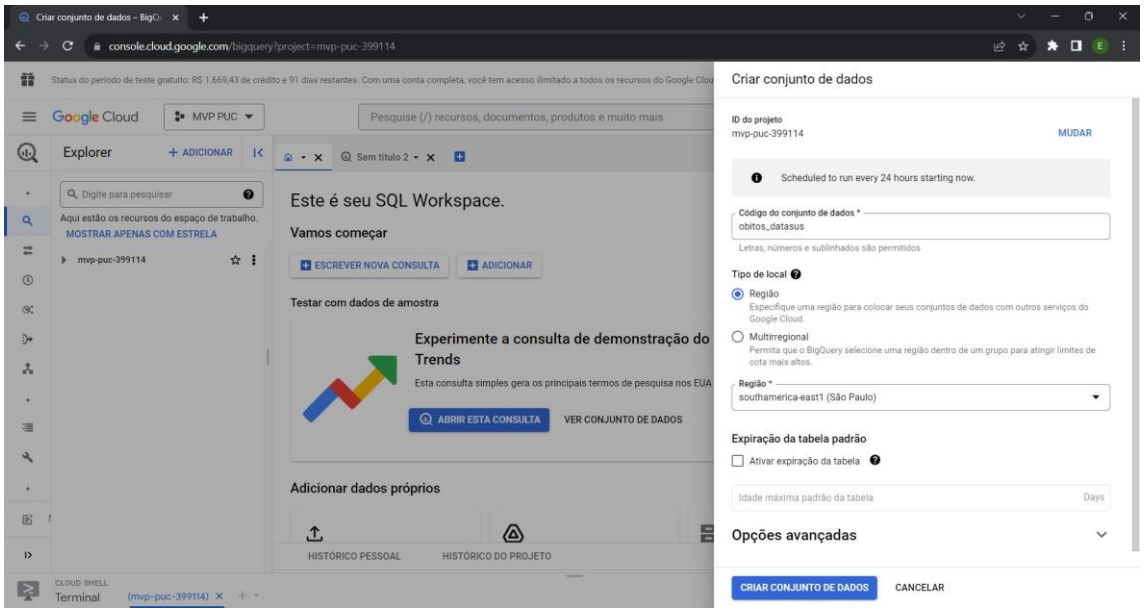


Figura 26: criação dataset no BigQuery

Por uma escolha para demonstração da ferramenta de consulta do BigQuery além de poder fazer a contabilização de todos os óbitos em relação ao

específico deste estudo, as tabelas geradas pelo Dataprep foram das causas externas com o enriquecimento pelas tabelas auxiliares, mas o filtro para suicídio ficou para ser feita através de consultas.

Contabilização de óbitos por causas externas por ano:

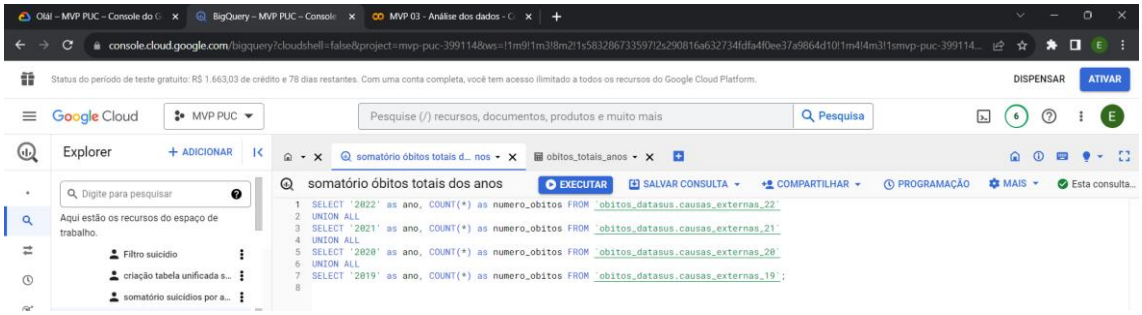


Figura 27: consulta contagem de óbitos causas externas

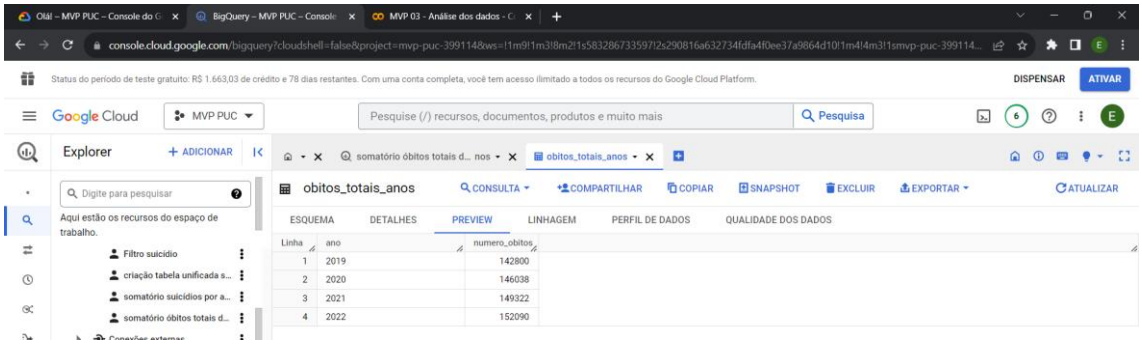


Figura 28: resultado consulta óbitos causas externas

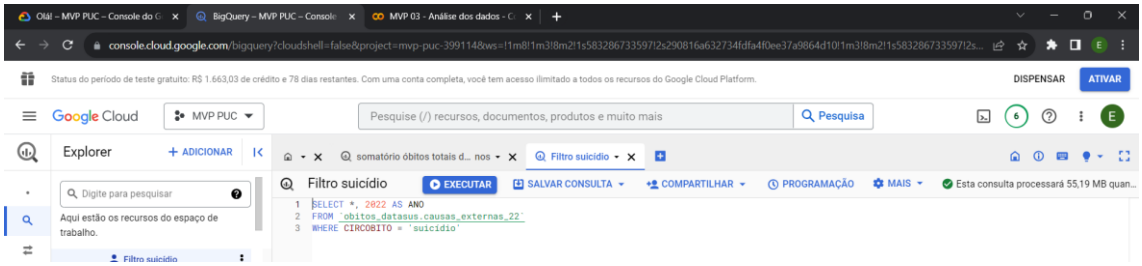


Figura 29: consulta com filtro para circunstância do óbito igual a suicídio

Destaca-se que cada consulta filtrada foi salva como nova tabela no próprio BigQuery porque em seguida seria feita a unificação em uma única tabela para fazer a análise exploratória. Em tempo, ressaltamos que na criação das tabelas de suicídio, foi a criação da coluna ANO para facilitar as manipulações futuras.

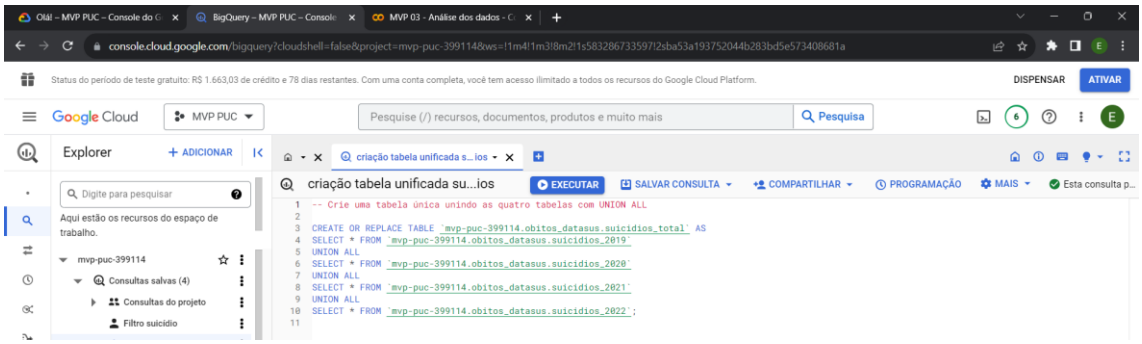


Figura 30: consulta para união das tabelas de suicídio

Por fim, a tabela unificada ficou acessível de maneira facilitada para o uso do Google Colab e do Looker Studio. No caso deste estudo, foi escolhido o Colab.

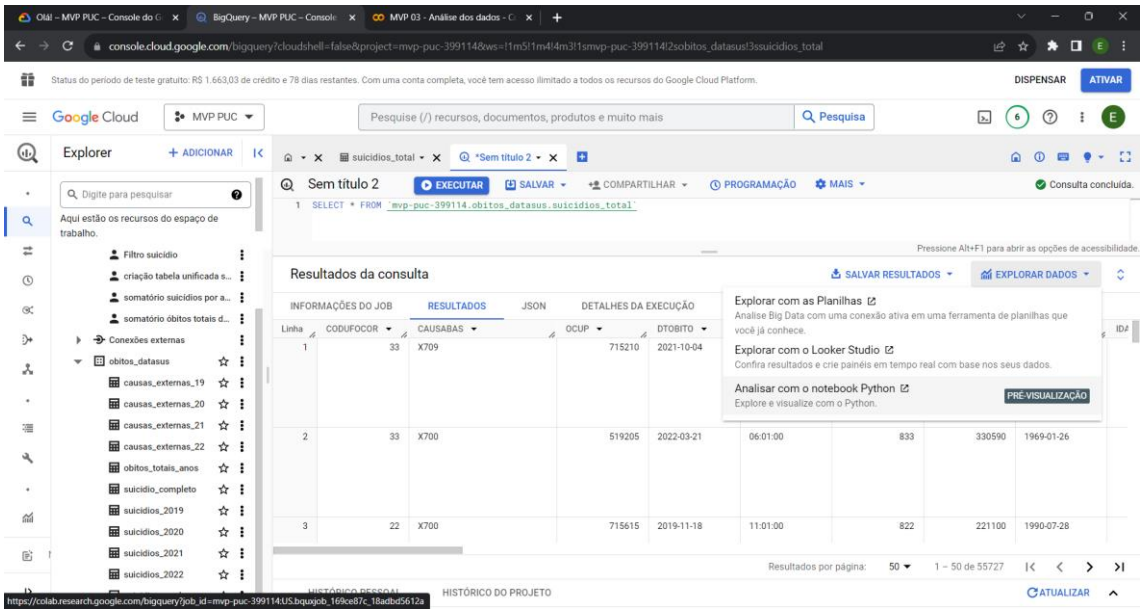


Figura 31: possibilidade de exploração dados do BigQuery com Colab

XIII. Análise

Como os dados de óbitos de causas externas do DATASUS é resultado da união de sistema cartorial e sistema de saúde, há muitas variáveis com dados faltantes, seja com a informação em branco, seja preenchida com o padrão adotado pelo DATASUS que é a palavra “ignorado”. De toda forma, essa realidade não inviabiliza o estudo baseado nestes dados.

Após o tratamento dos dados e exclusão de muitas colunas que não tem importância para o estudo proposto, a quantidade de valores faltantes é pequena e não trouxe problemas para a análise.

Um resumo da qualidade dos dados individualmente dos anos de 2019 a 2022 está em um relatório preparado pelo próprio Dataprep. Estes documentos estão disponibilizados ao final deste trabalho com os nomes de “Profile report – DOEXT19”, “Profile report – DOEXT20”, “Profile report – DOEXT21” e “Profile report – DOEXT22”.

Quanto à análise do perfil feita através do Google Colab, apresentar-se-á a seguir, mas destaca-se que a íntegra do notebook com códigos e resultados apresenta-se na parte dos anexos ou no Github: (https://github.com/dudu1626/MVP-Sprint-03/blob/bcdfa4c4951c4475757c41870cac8dc73db321bb/MVP_03_An%C3%A1lise_dos_dados_Com_modifica%C3%A7%C3%B5es.ipynb).

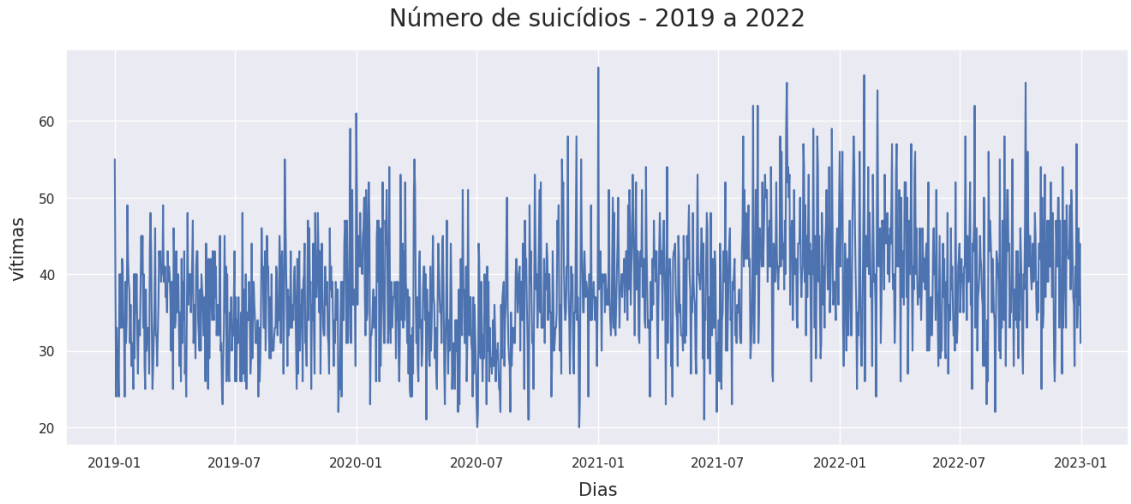
Resultados

1. Comparação entre os anos

ANO	NUM_SUICIDIOS	NUM_OBITOS	PERC_SUICIDIOS
2019	13.009	14.2800	9.11%
2020	13.166	14.6038	9.01%
2021	14.722	14.9322	9.86%
2022	14.830	15.2090	9.75%

Considerando o número de suicídios em relação ao número de mortes de causas externas totais, podemos perceber que houve um aumento de 9,1% em 2019 para 9,75% 2022. Em números absolutos o aumento foi de 1.821 suicídios.

2. Série temporal



A série temporal apresenta uma aparência de ruído aleatório, para conseguir fazer uma análise de tendência é mais conveniente o uso de média móvel.

Com a média móvel com 30 dias podemos perceber que o movimento respeita uma sazonalidade interessante, podemos perceber que há queda nos casos nos meses de julho e crescimento no final do ano. Corroborando a análise anterior, observa-se que os casos estão subindo da média de 34 por mês em janeiro de 2019 para mais de 40 em dezembro de 2022.

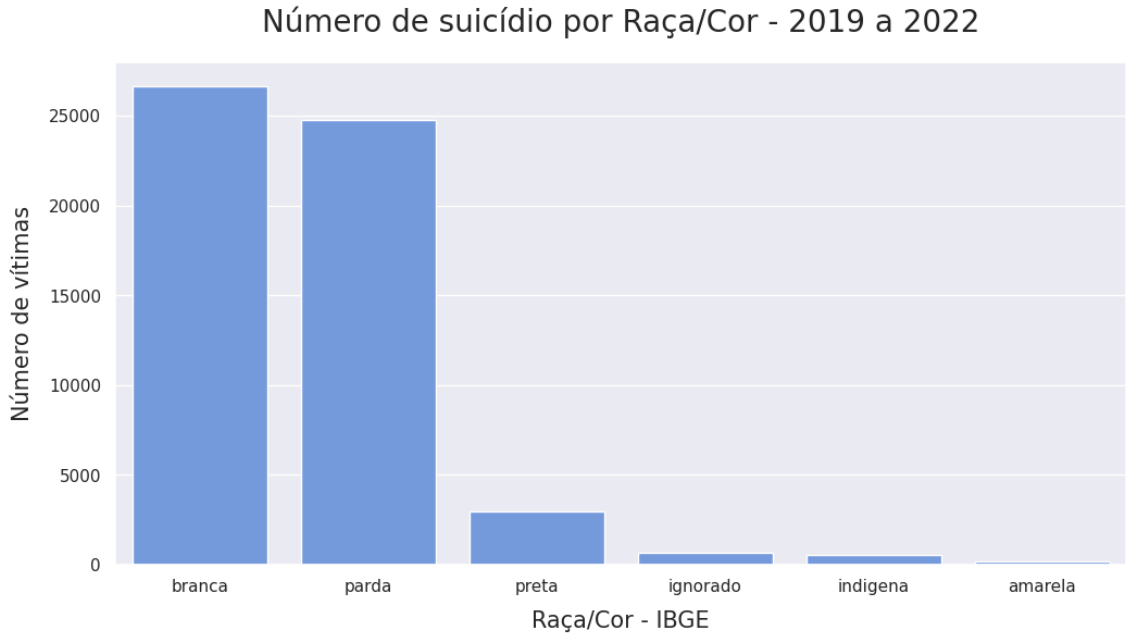
Uma oportunidade de pesquisa seria descobrir as possíveis causas que causam esse fenômeno na sazonalidade.

3. Análise das variáveis

A análise das variáveis foi feita considerando os 4 anos juntos e depois separados ano a ano para evitar erro de julgamento devido ao Paradoxo de Simpson (https://pt.wikipedia.org/wiki/Paradoxo_de_Simpson)

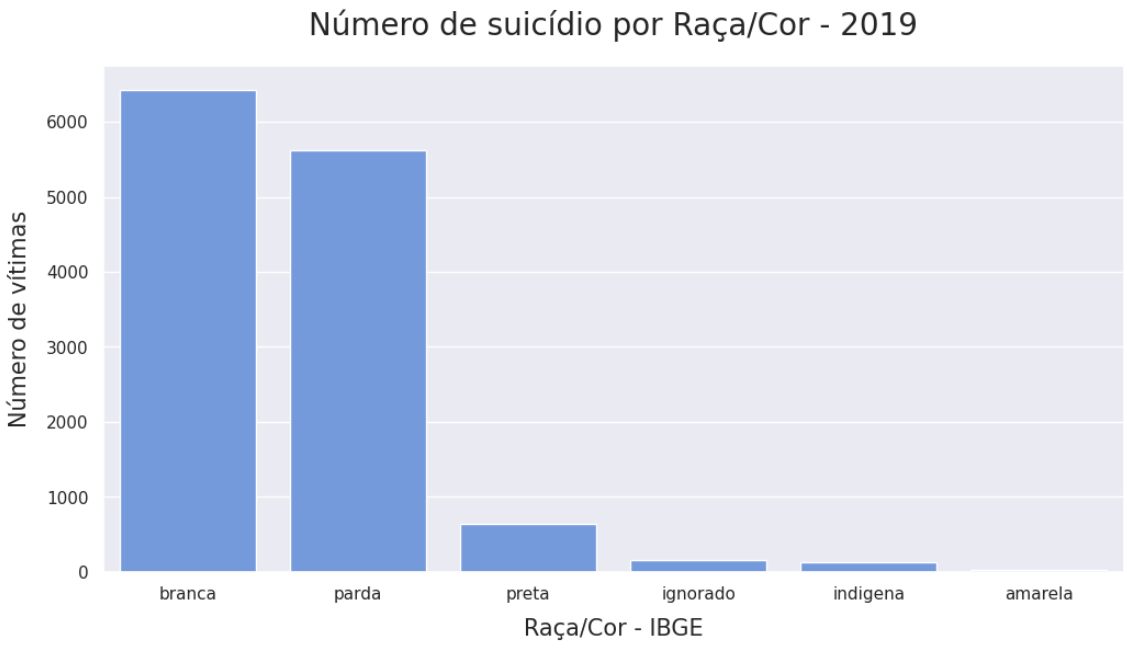
a) Suicídio por raça/cor

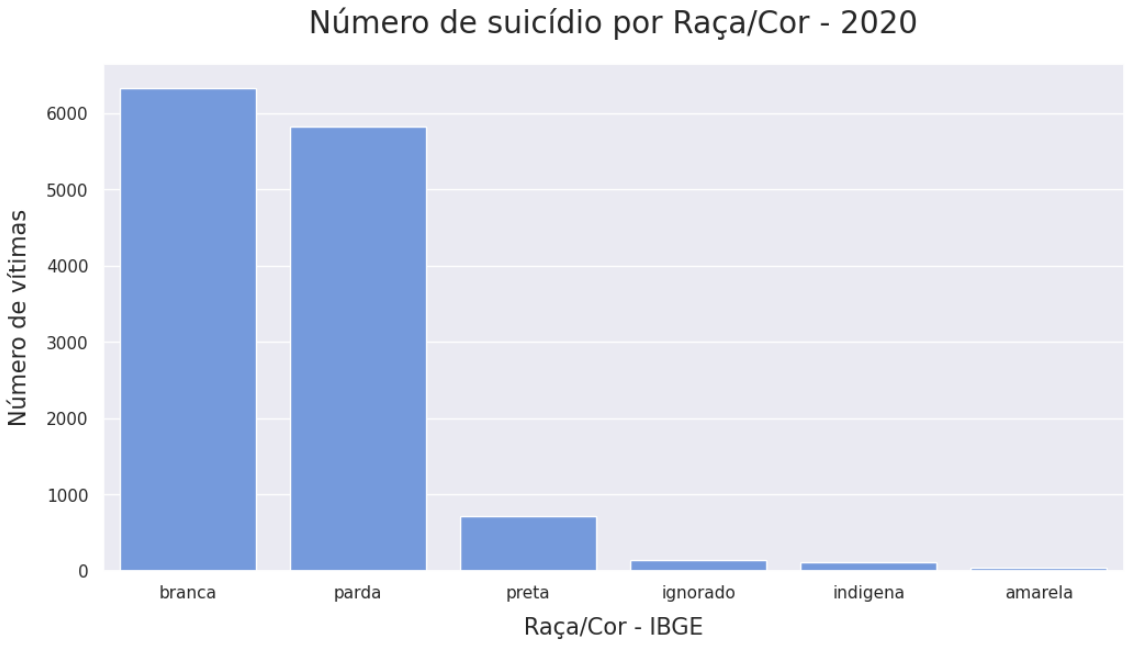
Raça/Cor	Percentual
branca	47.81%
parda	44.40%
preta	5.32%
ignorado	1.23%
indígena	0.94%
amarela	0.20%

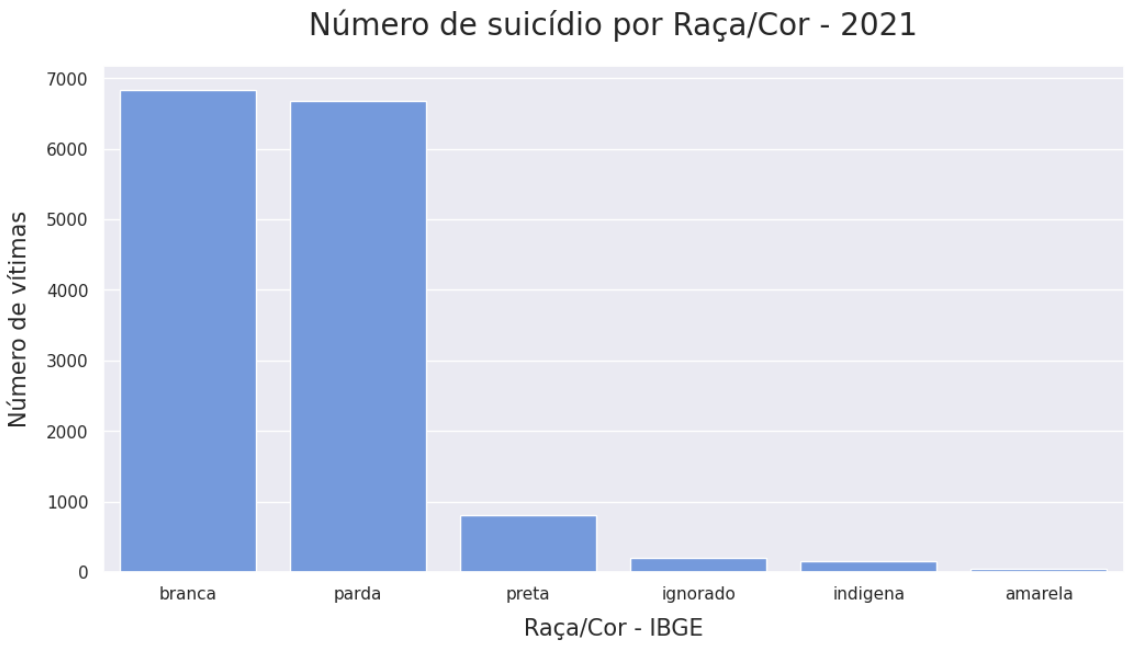


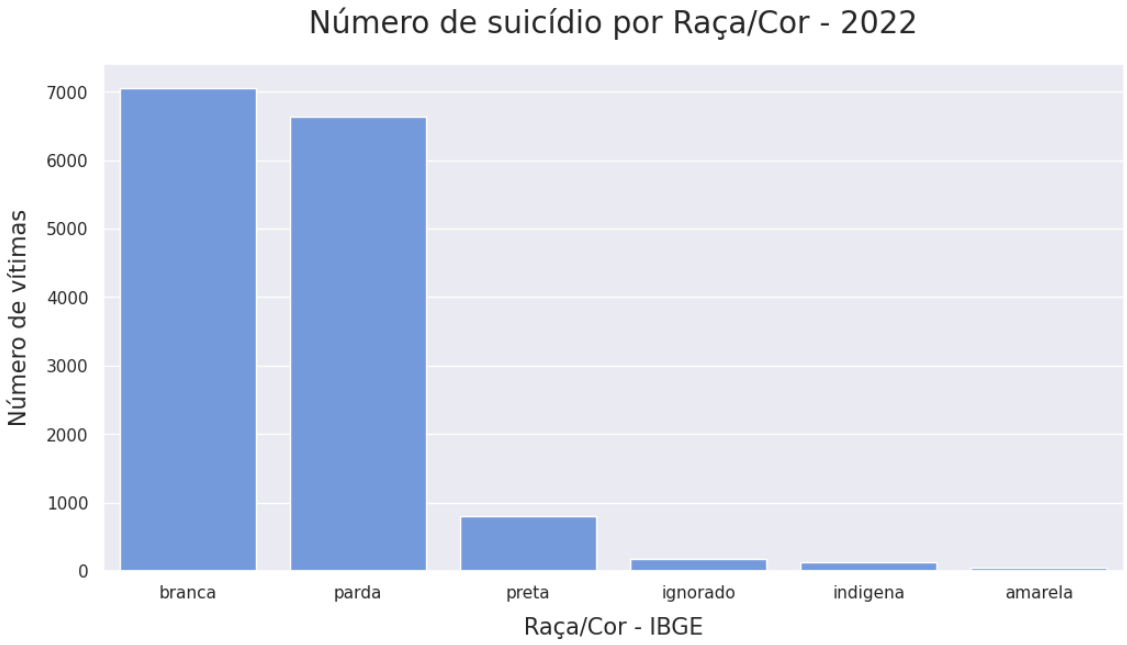
O que podemos observar é que quase metade dos suicídios foram de brancos (47,8%), seguidos de perto pelos pardos (44,4%), cabendo ressaltar que a incidência para pretos é proporcionalmente baixa (5,3%).

Quanto a distribuição de casos de ano a ano, podemos observar que respeita a mesma regra do somatório dos 4 anos:


Número de suicídio por Raça/Cor - 2019
Número de vítimas
6000
5000
4000
3000
2000
1000
0
branca
parda
preta
ignorado
indigena
amarela
Raça/Cor - IBGE

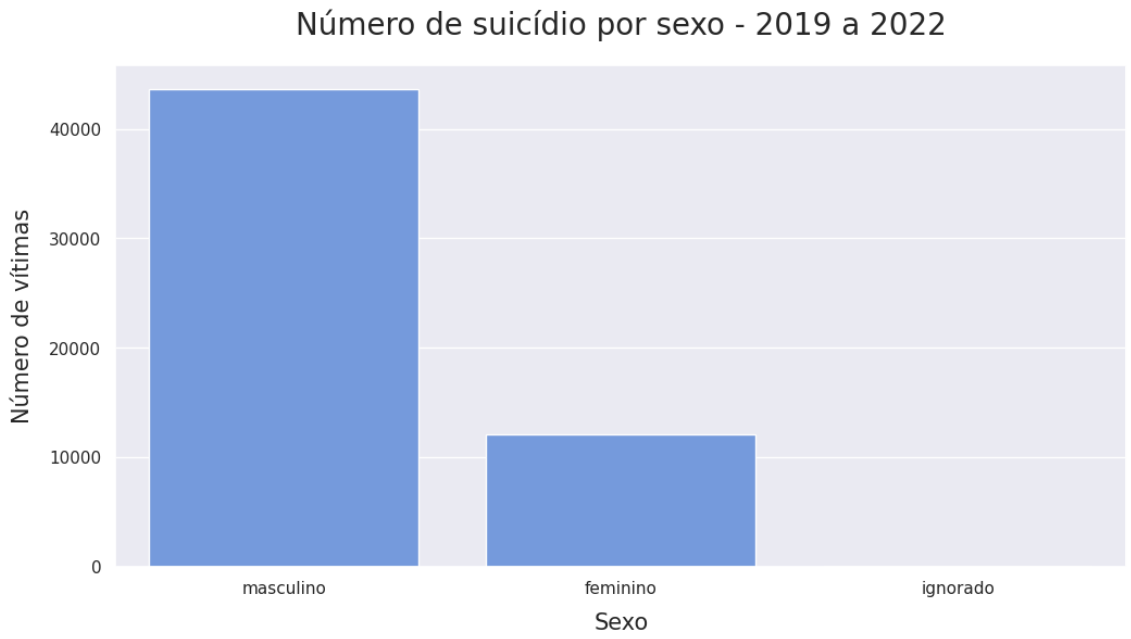

Número de suicídio por Raça/Cor - 2020
Número de vítimas
6000
5000
4000
3000
2000
1000
0
branca
parda
preta
ignorado
indigena
amarela
Raça/Cor - IBGE





b) Suicídio por sexo

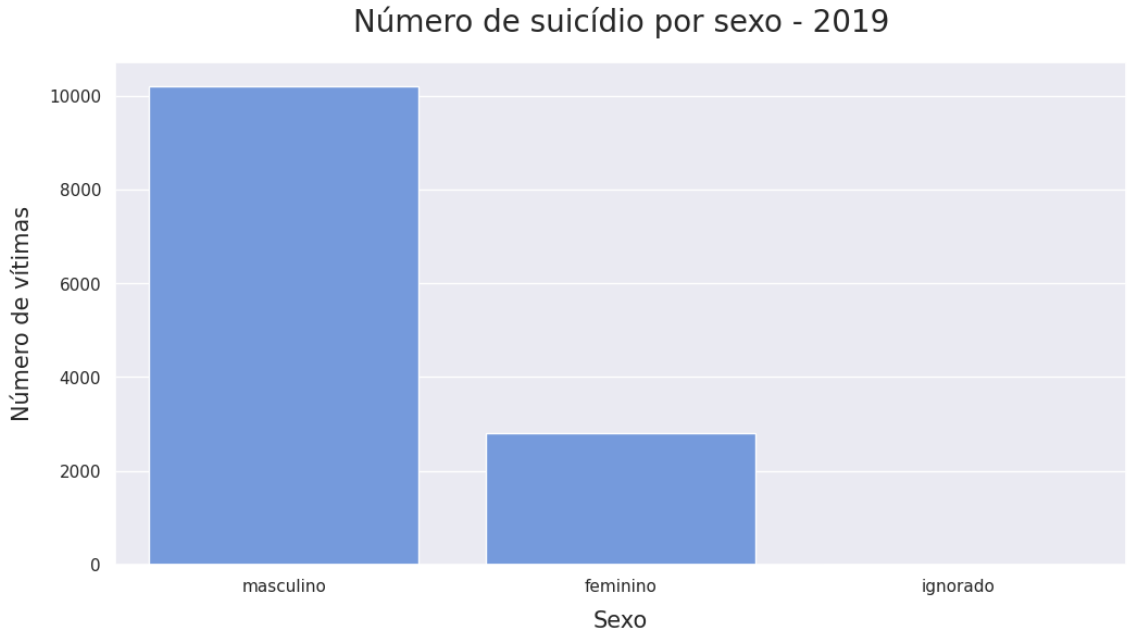
Sexo	Percentual
Masculino	78.20%
Feminino	21.68%
Ignorado	0.02%

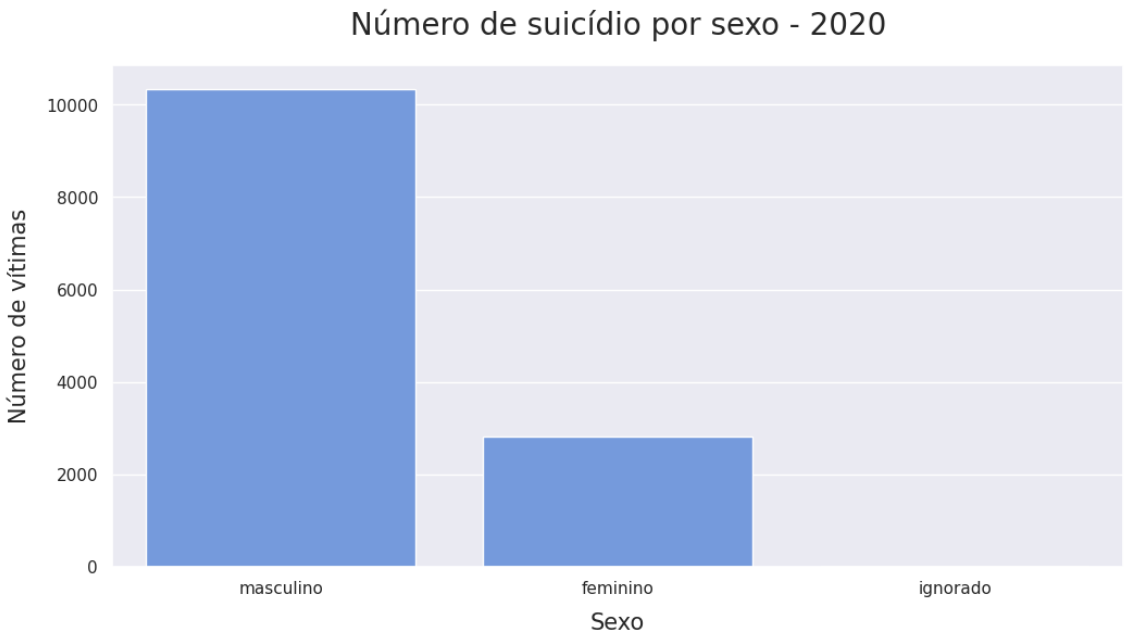


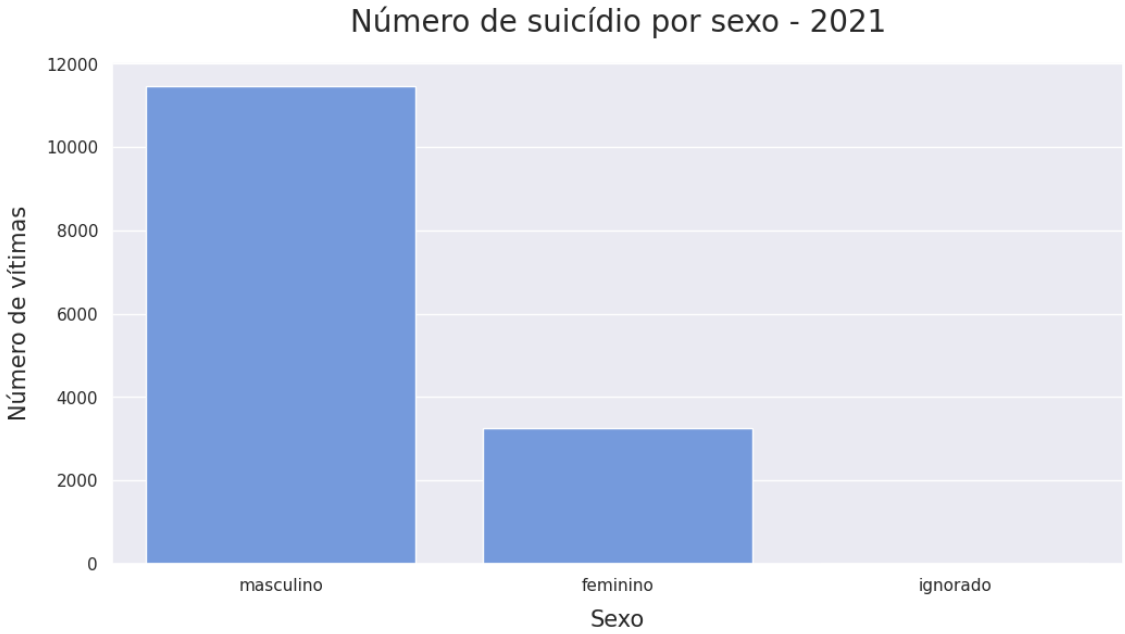
Podemos observar que a diferença entre homens e mulheres é bastante relevante. Enquanto os homens têm mais de 40 mil casos (78%), é pouco mais de 10 mil o caso de mulheres (21,68%).

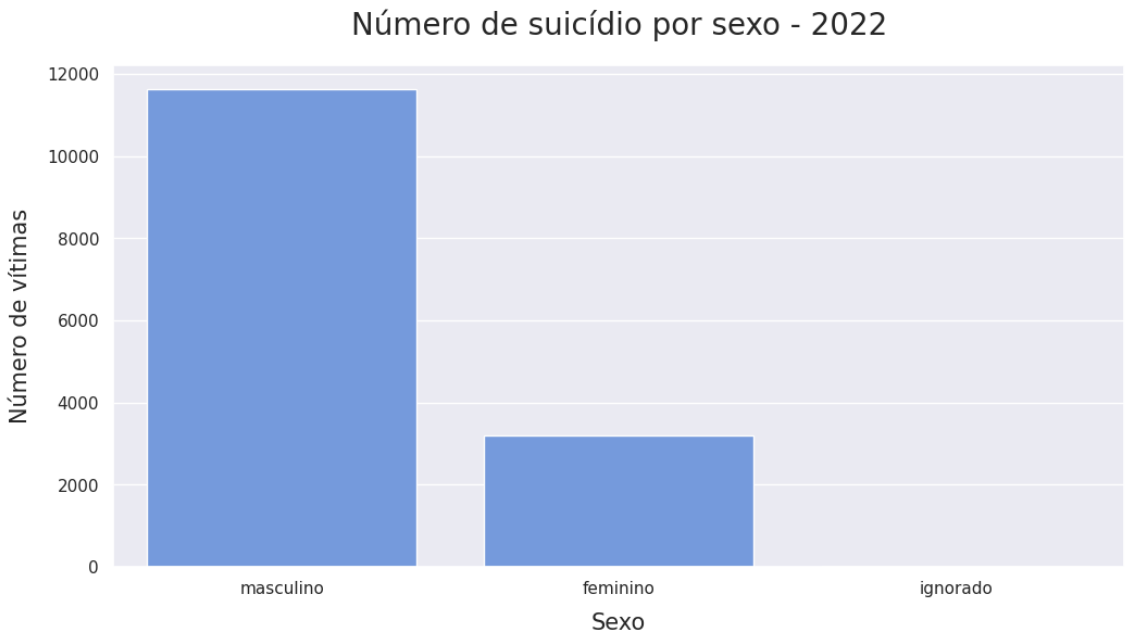
Infelizmente não existe a possibilidade de fazer análise quanto às identidades de gêneros.

Da mesma forma que a raça/cor, a distribuição de casos de ano a ano, respeita a mesma regra do somatório dos 4 anos:




Número de suicídio por sexo - 2020
Número de vítimas
10000
8000
6000
4000
2000
0
masculino
feminino
ignorado
Sexo

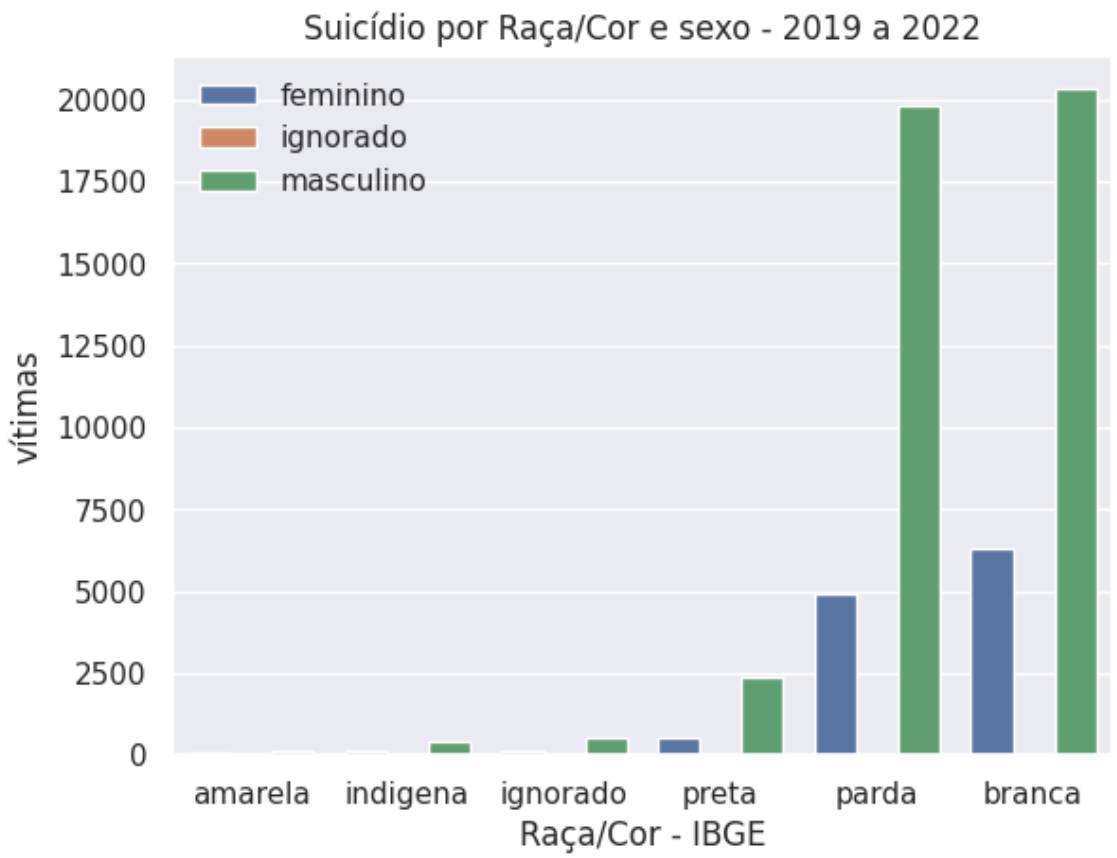

Número de suicídio por sexo - 2021
Número de vítimas
12000
10000
8000
6000
4000
2000
0
masculino
feminino
ignorado
Sexo



Mesmo observando a superioridade numérica entre homens e mulheres, somente pode-se afirmar que os homens são mais efetivos quando se propõe a tirar a própria vida, caberia um estudo futuro, com uma base de dados com as informações acerca da tentativa para ter certeza se essa diferença permanece e gerar política de mitigação desses casos focada nos homens.

c) Cruzamento entre raça e sexo

RACACOR	amarela	branca	ignorado	indígena	parda	preta
SEXO						
feminino	45	6.323	128	115	4.911	562
ignorado	0	0	7	0	2	1
masculino	122	20.319	548	411	19.830	2.403



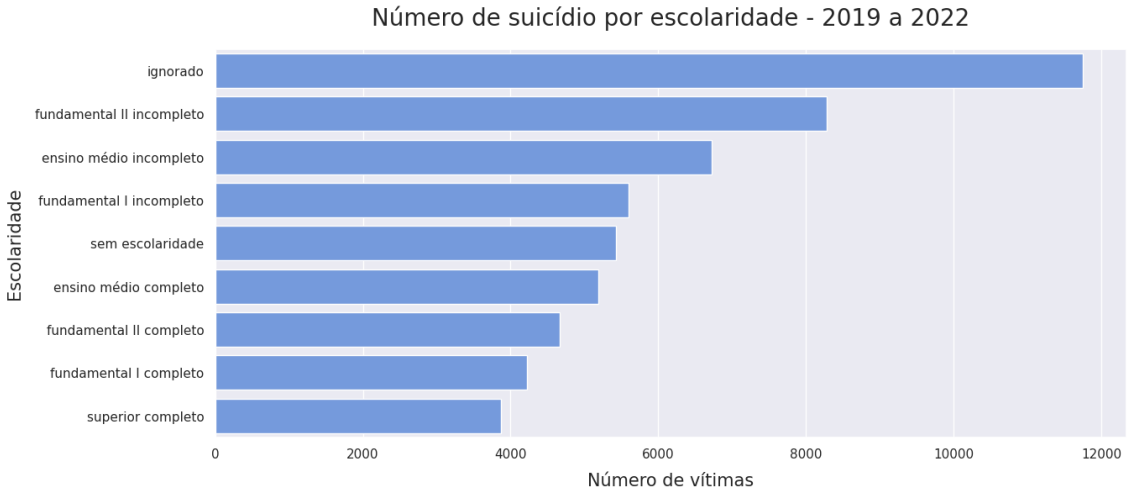
Não há diferença perceptível quando se separa a raça/cor pelo sexo para os casos de suicídios.

d) Suicídio por escolaridade

Inicialmente cabe informar que a metodologia de contabilização de escolaridade é confusa. São três variáveis destinadas a este fim, mas apresentam números inconsistentes entre si, contudo as variáveis ESC2010 e ESCFALAGR1 têm valores comuns e dentre eles foi escolhida a coluna ESCFALAGR1, pois apresenta maior detalhamento.

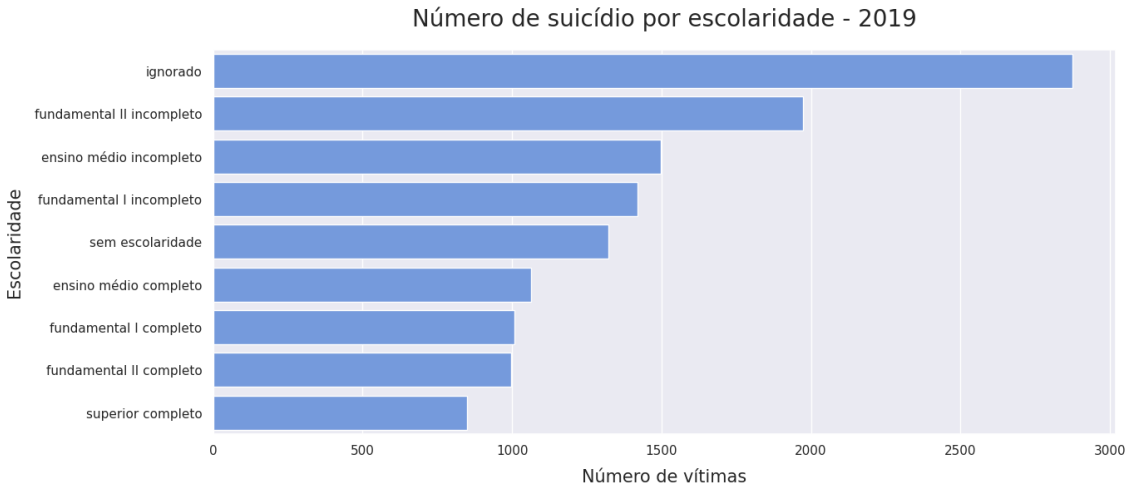
Escolaridade	Percentual
Ignorado	21.09%
fundamental II incompleto	14.87%
ensino médio incompleto	12.07%
fundamental I incompleto	10.04%
sem escolaridade	9.74%
ensino médio completo	9.30%

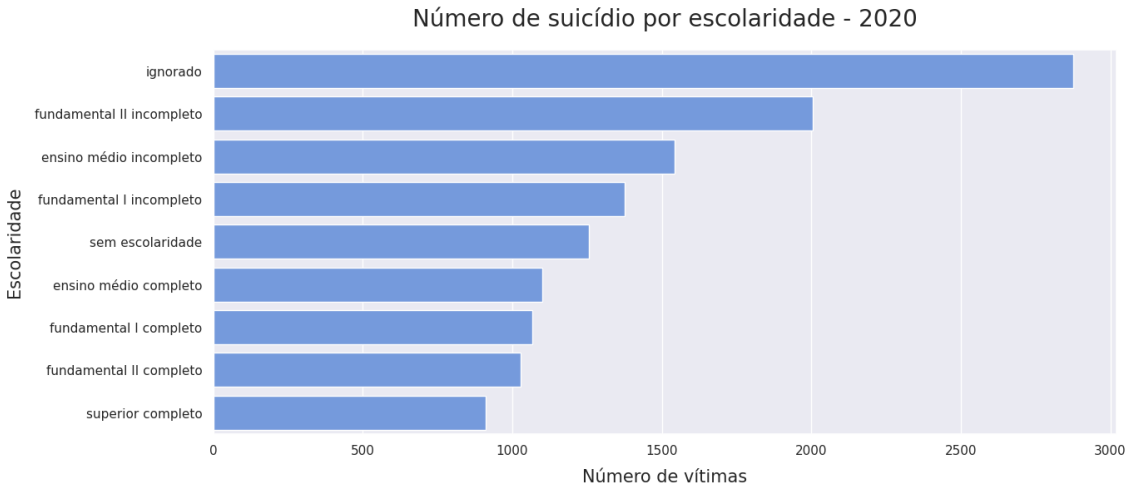
fundamental II completo	8.37%
fundamental I completo	7.58%
superior completo	6.95%

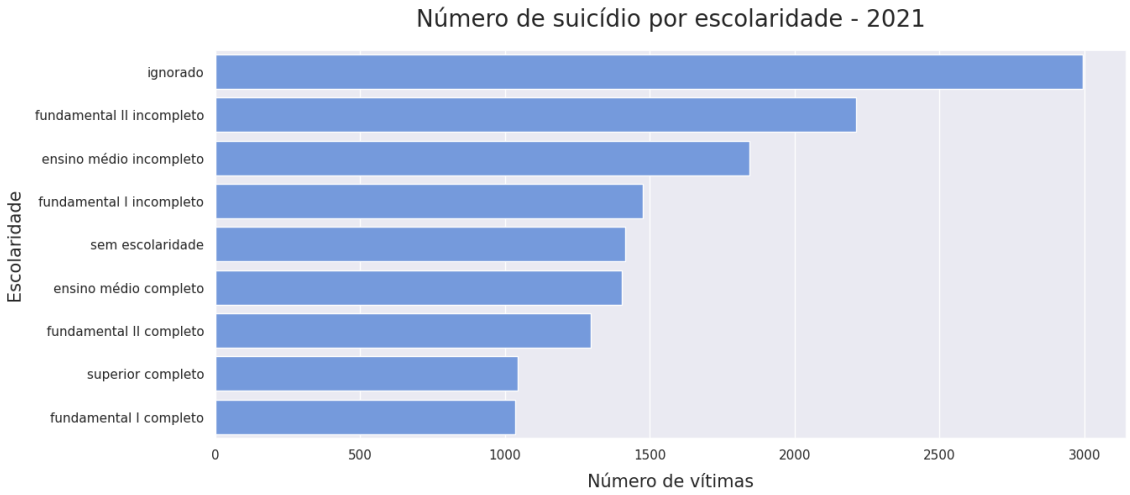


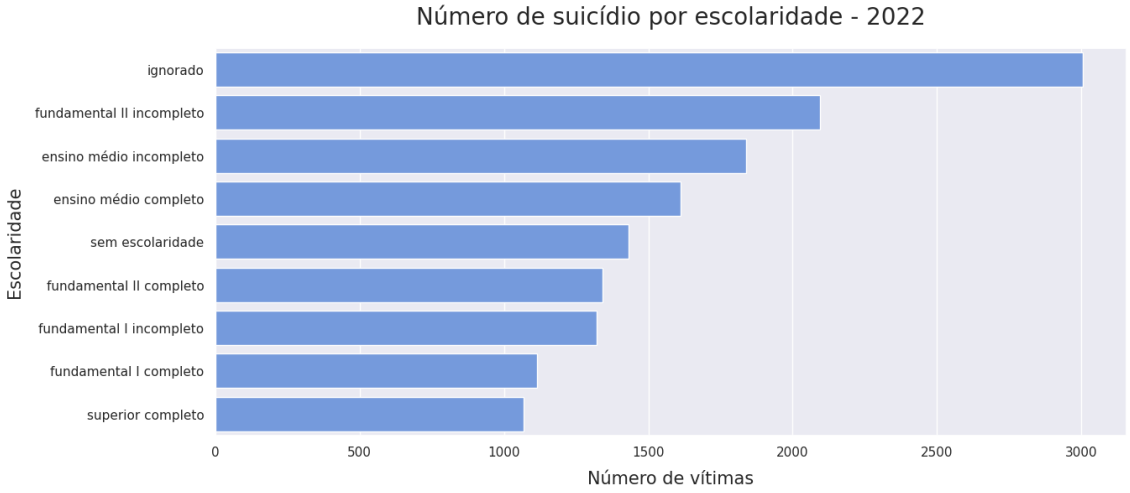
Excluindo a parcela de escolaridade ignorada, Podemos observar que a distribuição de suicídios é maior na parcela da população com menor escolaridade.

A seguir, observaremos que não há diferença se observar os anos separadamente.









e) Comparação dos suicídios, comparando sexo e escolaridade

ESCOLARIDADE	ensino médio completo	ensino médio incompleto	fundamental I completo	fundamental I incompleto	fundamental II completo	fundamental II incompleto	ignorado	sem escolaridade	superior completo
SEXO									
feminino	1294	1724	732	902	939	1607	2300	1285	1301
ignorado	0	1	0	0	0	0	9	0	0
masculino	3890	5000	3492	4691	3727	6678	9442	4142	2571

Há de se destacar que a diferença entre homens e mulheres é de, no mínimo de 200% (ensino médio completo) em quase todas as escolaridades, exceto com o nível superior completo que é de 97%.

f) Suicídio por hora

A escolha da hora do óbito é um dado peculiar porque pode-se observar que há uma preferência pelo horário do dia em detrimento da noite ou madrugada.

Considerando os 4 anos analisados, o horário das 18h é o primeiro horário noturno, que se encontra no 5º lugar e depois deste é às 19h em 13º lugar. Se considerarmos os anos separadamente, percebemos o mesmo fenômeno, com pequenas variações.

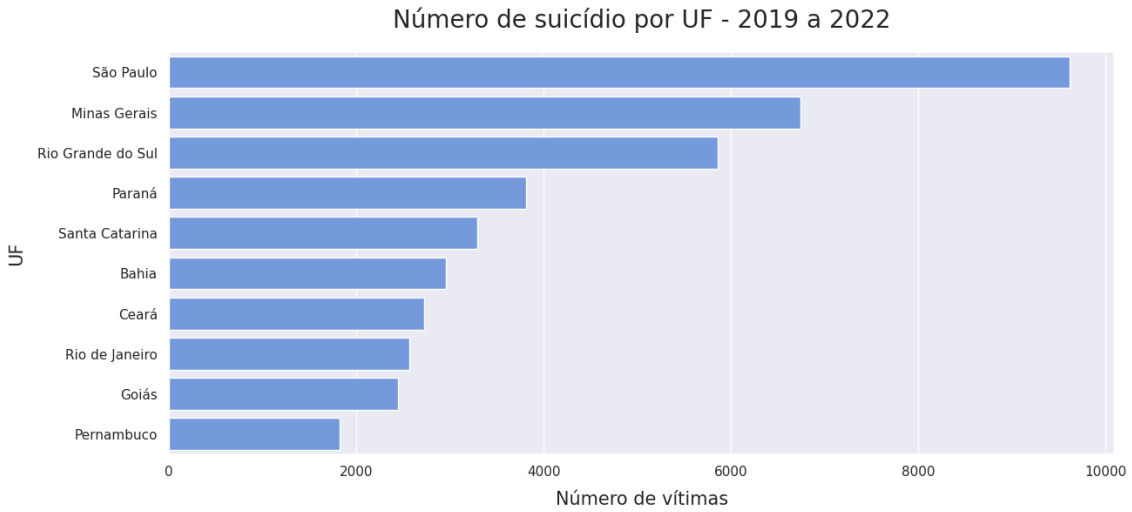
g) Suicídio por Unidade Federativa (Estado)

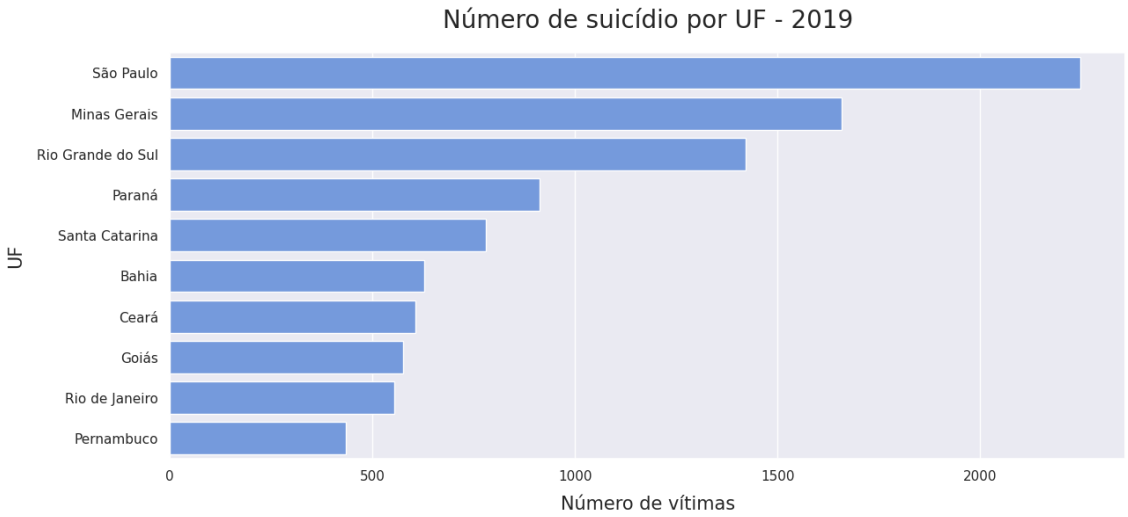
Se considerarmos os números brutos nos 4 anos somados encontraremos o cenário conforme quadro a seguir, destacando que os três estados da Região Sul estão nos primeiros lugares, mesmo sabendo que não são dos mais populosos e ressalta-se ainda que os 5 primeiros estados correspondem a mais de 50% de todos os casos (52.60%):

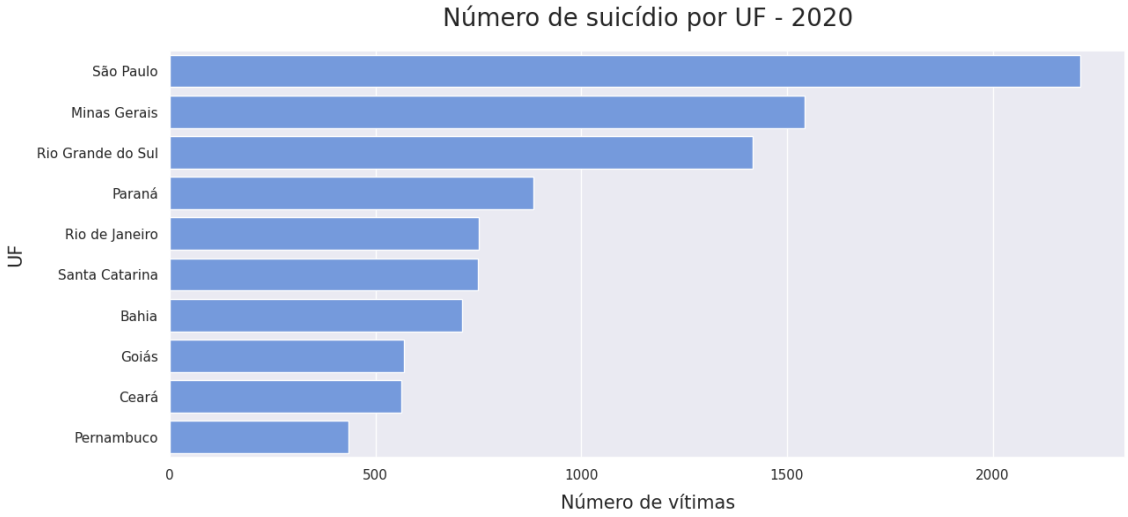
UF	contagem	percentual
São Paulo	9.614	17.25%
Minas Gerais	6.739	12.09%
Rio Grande do Sul	5.855	10.51%
Paraná	3.815	6.85%
Santa Catarina	3.291	5.91%
Bahia	2.959	5.31%
Ceará	2.725	4.89%
Rio de Janeiro	2.566	4.60%
Goiás	2.449	4.39%
Pernambuco	1.824	3.27%
Pará	1.550	2.78%
Piauí	1.342	2.41%
Maranhão	1.265	2.27%
Amazonas	1.144	2.05%
Mato Grosso do Sul	1.106	1.98%
Paraíba	1.052	1.89%
Mato Grosso	985	1.77%
Espírito Santo	939	1.68%
Rio Grande do Norte	871	1.56%
Distrito Federal	788	1.41%
Alagoas	591	1.06%
Rondônia	556	1.00%

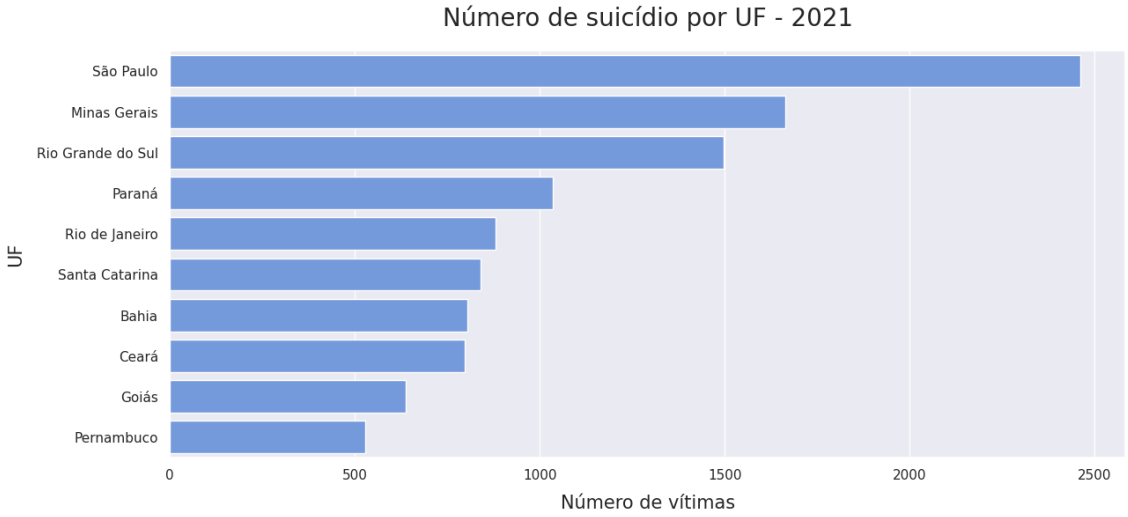
Tocantins	541	0.97%
Sergipe	526	0.94%
Acre	273	0.49%
Amapá	213	0.38%
Roraima	148	0.26%

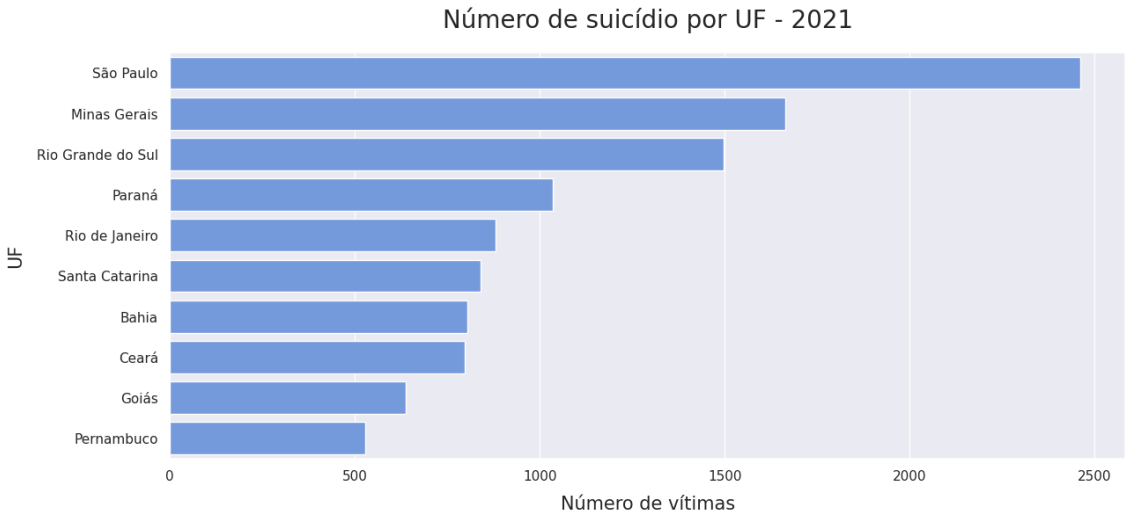
Para a melhor apresentação, apresentar-se-á apenas os 10 estados com a maior prevalência.




Número de suicídio por UF - 2019
São Paulo
Minas Gerais
Rio Grande do Sul
Paraná
Santa Catarina
Bahia
Ceará
Goiás
Rio de Janeiro
Pernambuco
UF
0
500
1000
1500
2000
Número de vítimas


Número de suicídio por UF - 2020
São Paulo
Minas Gerais
Rio Grande do Sul
Paraná
Rio de Janeiro
Santa Catarina
Bahia
Goiás
Ceará
Pernambuco
UF
0
500
1000
1500
2000
Número de vítimas


Número de suicídio por UF - 2021
São Paulo
Minas Gerais
Rio Grande do Sul
Paraná
Rio de Janeiro
Santa Catarina
Bahia
Ceará
Goiás
Pernambuco
UF
0
500
1000
1500
2000
2500
Número de vítimas



Para o ano de 2022, podemos comparar com os dados da população informada pelo IBGE, em consequência pode-se alterar o foco da análise e, consequentemente, a situação avaliada em alguns estados.

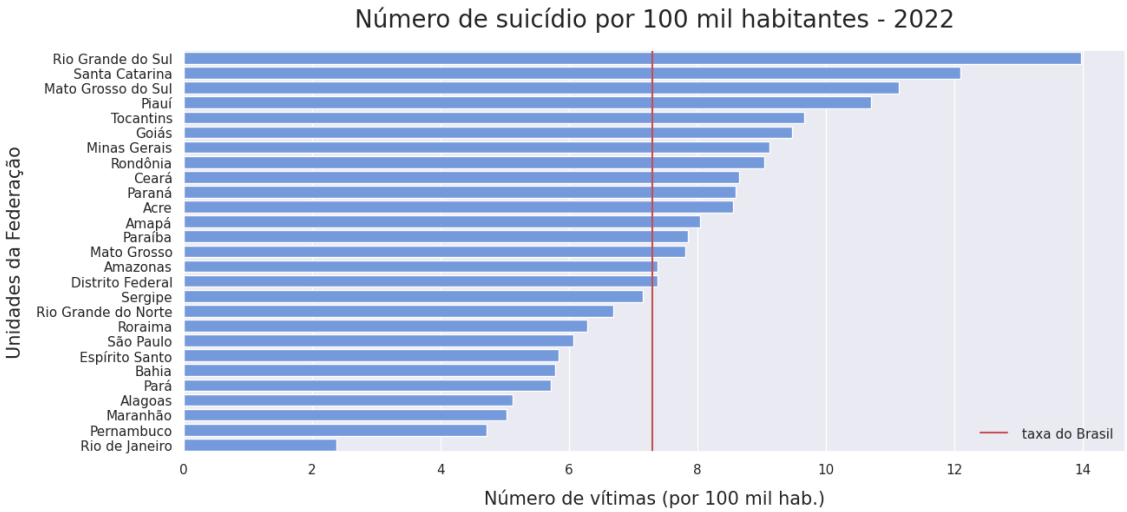
Os estados de São Paulo, Pernambuco, Bahia e Rio de Janeiro por serem estados populosos, ficavam entre os 10 maiores em números absolutos, mas se considerar a taxa de suicídios por cem mil habitantes eles ficam entre os 10 menores casos, com destaque para o Rio de Janeiro que fica em último lugar com uma taxa metade do penúltimo.

O destaque negativo fica para o Estado do Rio Grande do Sul que é o terceiro estado com o maior número de suicídios em números absolutos e o primeiro a cada 100 mil habitantes.

Uma informação importante para comparação é a taxa de suicídio do Brasil em 2022 é de 7,30 por cada 100 mil habitantes.

UF	contagem	População 2022	suicídio por cem mil habitantes
Rio Grande do Sul	1520	10880506	13.97
Santa Catarina	921	7609601	12.10
Mato Grosso do Sul	307	2756700	11.14
Piauí	350	3269200	10.71
Tocantins	146	1511459	9.66
Goiás	668	7055228	9.47

Minas Gerais	1873	20538718	9.12
Rondônia	143	1581016	9.04
Ceará	760	8791688	8.64
Paraná	984	11443208	8.60
Acre	71	830026	8.55
Amapá	59	733508	8.04
Paraíba	312	3974495	7.85
Mato Grosso	286	3658813	7.82
Amazonas	291	3941175	7.38
Distrito Federal	208	2817068	7.38
Sergipe	158	2209558	7.15
Rio Grande do Norte	221	3302406	6.69
Roraima	40	636303	6.27
São Paulo	2696	44420459	6.07
Espírito Santo	224	3833486	5.84
Bahia	818	14136417	5.79
Pará	464	8116132	5.72
Alagoas	160	3127511	5.12
Maranhão	341	6775152	5.03
Pernambuco	427	9058155	4.71
Rio de Janeiro	382	16054524	2.38



h) Suicídio por ocupação

Inicialmente, é importante destacar que essa variável apresenta 41% classificado como ignorado, mas entende-se como uma informação de muita importância, então cortamos esse dado para poder avaliar o perfil das pessoas com essa informação cadastrada.

Considerando a ocupação (trabalho que desempenha) das pessoas que se suicidam, podemos observar uma informação que corrobora o que se observou na avaliação por escolaridade, uma vez que as ocupações que se apresentam são profissões que exigem baixa formação:

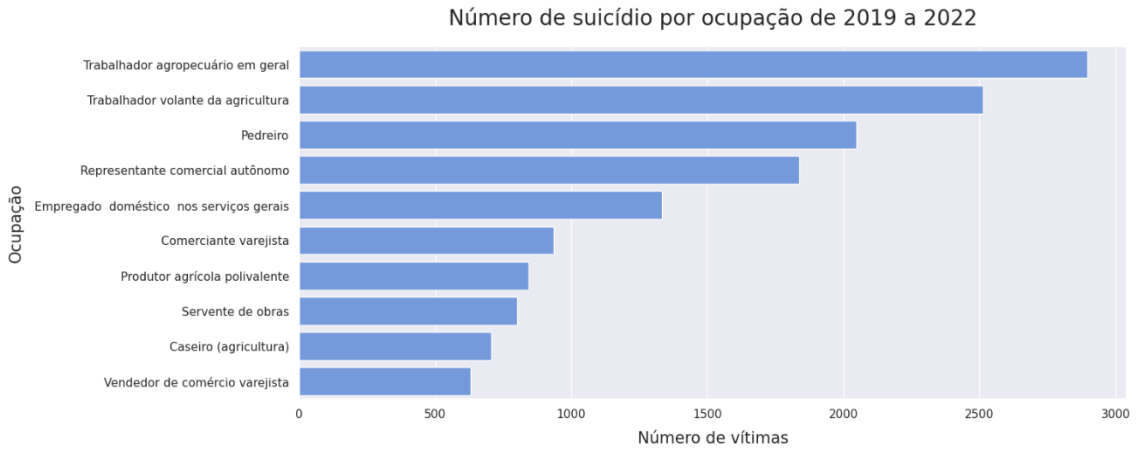
A seguir apresenta-se a lista das 20 ocupações com os maiores números de suicídio dos anos de 2019 a 2022 somados. A informação percentual apresentada a seguir é em relação às pessoas que foram classificadas com alguma ocupação.

Ocupação	Contagem	Percentual
Trabalhador agropecuário em geral	2896	8.98%
Trabalhador volante da agricultura	2512	7.79%
Pedreiro	2047	6.34%
Representante comercial autônomo	1837	5.69%
Empregado doméstico nos serviços gerais	1334	4.13%
Comerciante varejista	935	2.90%
Produtor agrícola polivalente	843	2.61%
Servente de obras	802	2.49%
Caseiro (agricultura)	705	2.18%
Vendedor de comércio varejista	631	1.96%
Motorista de carro de passeio	575	1.78%
Pintor de obras	419	1.30%
Mecânico de manutenção de automóveis, motocicletas e veículos similares	376	1.16%
Vigilante	374	1.16%
Motorista de caminhão (rotas regionais e internacionais)	356	1.10%

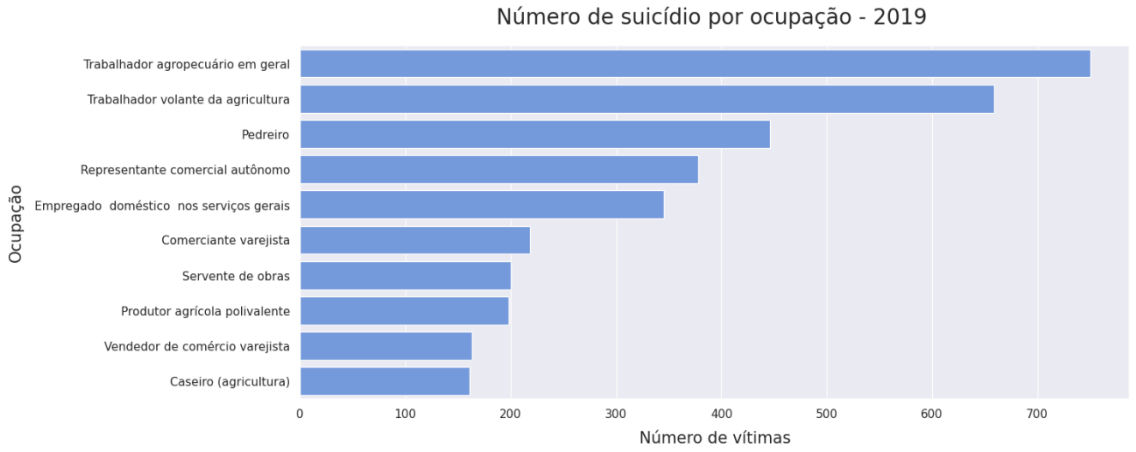
Alimentador de linha de produção	308	0.95%
Advogado	295	0.91%
Assistente administrativo	280	0.86%
Técnico de enfermagem	273	0.85%
Eletricista de instalações	242	0.75%

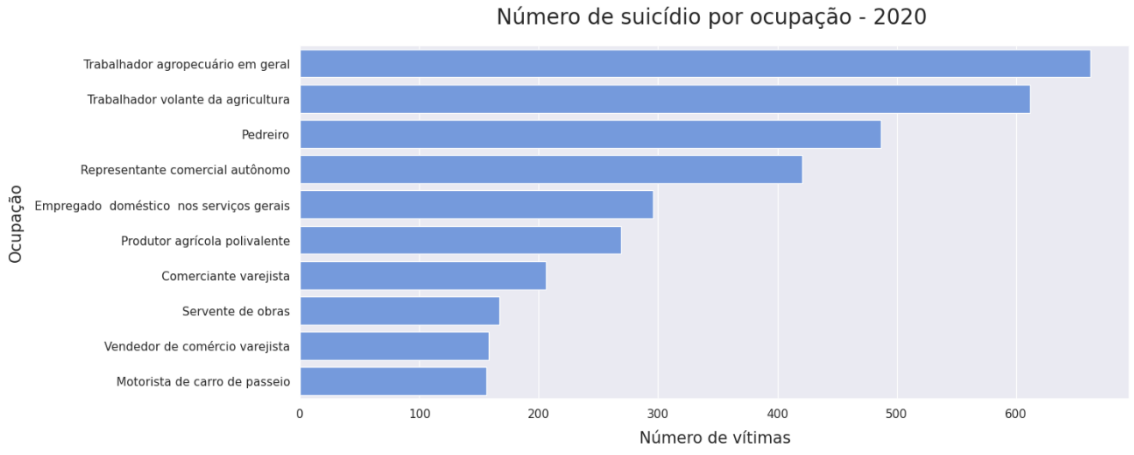
Uma ocupação que foge à regra da baixa escolaridade é a de advogado. Uma oportunidade de estudo para identificar as causas que faz pessoas com esta profissão tomarem esta atitude.

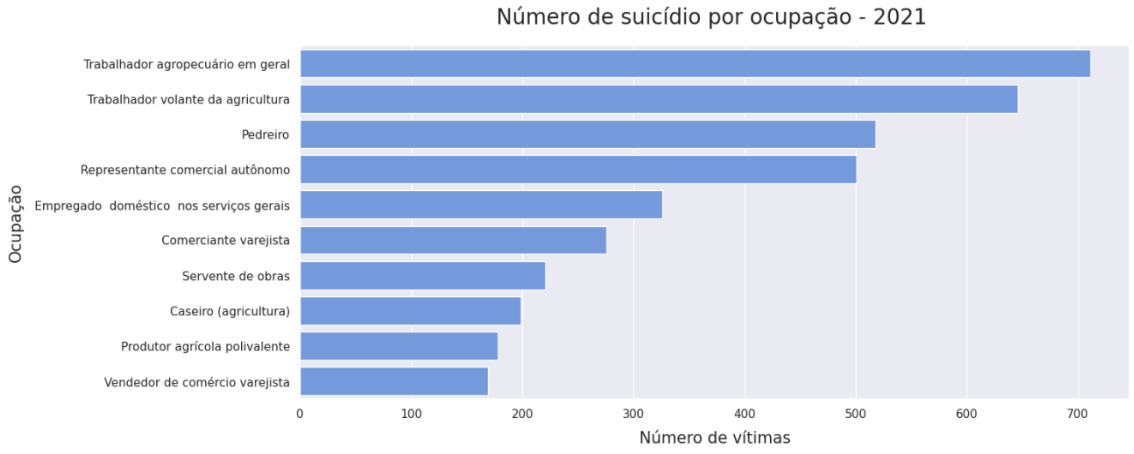
Abaixo podemos observar o gráfico com as 10 primeiras ocupações:

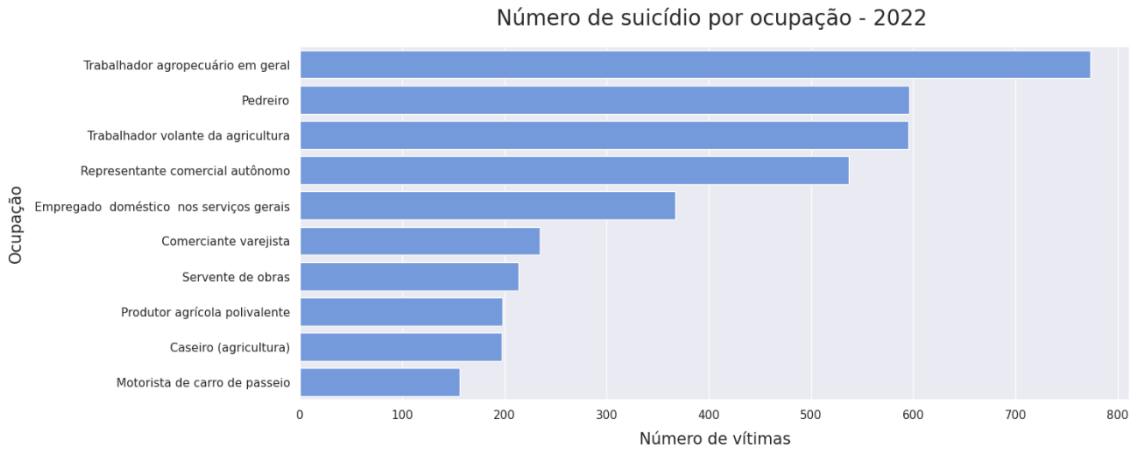


Podemos observar que essas profissões permanecem conforme os anos.









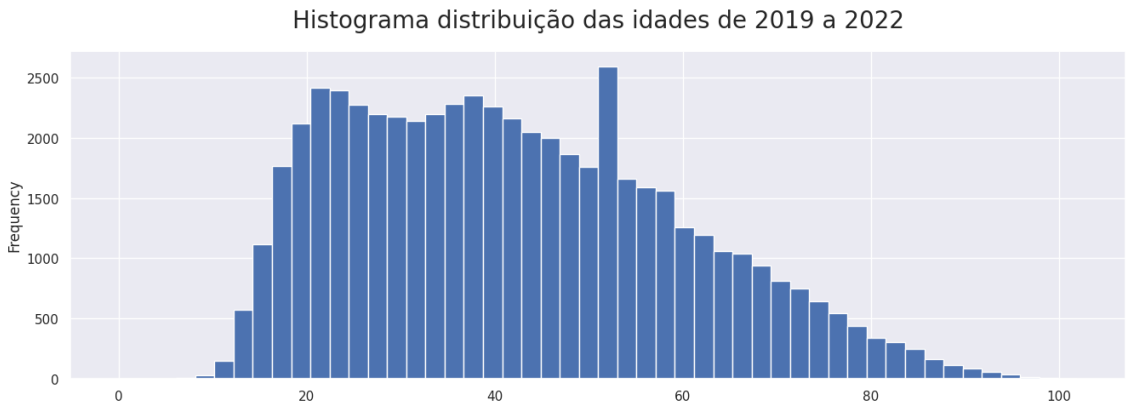
i) Suicídio por idade

Quanto a análise de suicídio por idade, podemos identificar que alguns valores são incompatíveis com a realidade, o que nos faz acreditar que houve problema no preenchimento das informações, como exemplo a idade mínima identificada é de zero anos.

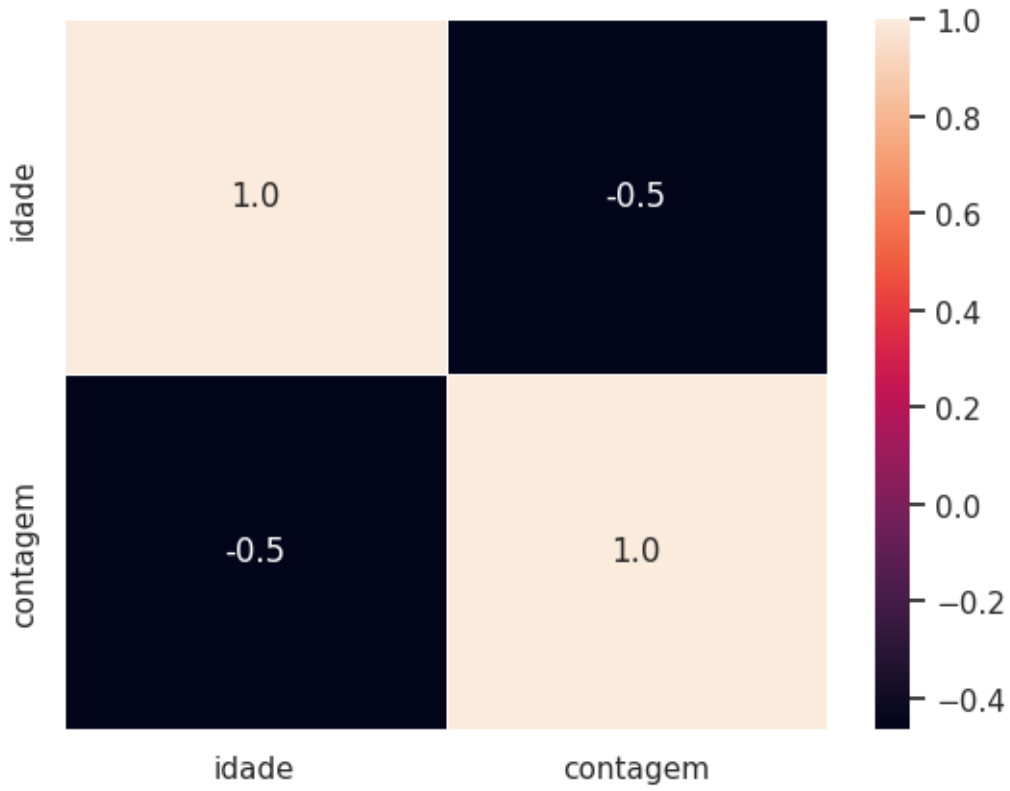
Resumo estatístico de 2019 a 2022

Medidas estatísticas	Idade
contagem	55.727
média	42.068
Desvio padrão	17.629
mín.	0.0
25%	27
50%	40
75%	55
máx.	102

A distribuição das idades é conforme o histograma:



Embora a distribuição tenha uma aparência de distribuição normal, tentou-se um teste de correlação, com o intuito de ter algum insight, mas não foi identificada a correlação.



As distribuições de suicídio por idade separadamente por ano não apresentaram diferença que se destaque.

Resumo estatístico de 2019

Medidas estatísticas	Idade
contagem	13009
média	41.537
Desvio padrão	17.49
mín.	5
25%	27
50%	39
75%	54
máx.	102

Resumo estatístico de 2020

Medidas estatísticas	Idade
contagem	13166
média	42.590

Desvio padrão	17.878
mín.	8
25%	28
50%	41
75%	55
máx.	102

Resumo estatístico de 2021

Medidas estatísticas	Idade
contagem	14722
média	42.188
Desvio padrão	17.773
mín.	0
25%	27
50%	40
75%	55
máx.	99

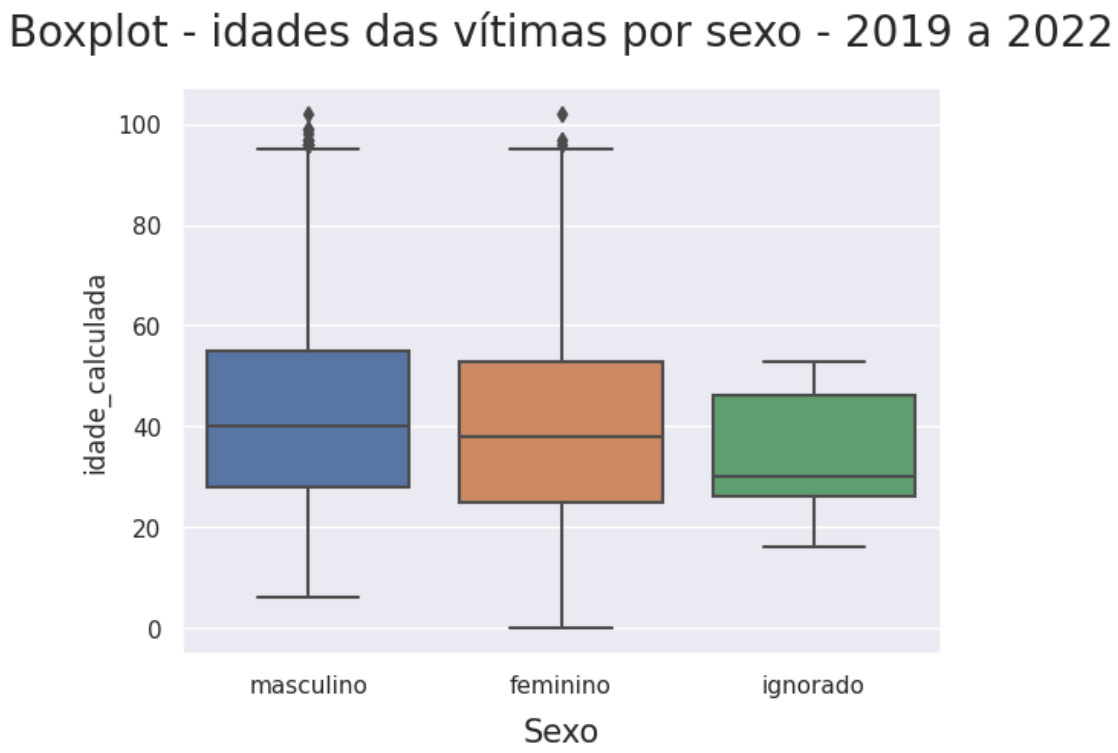
Resumo estatístico de 2022

Medidas estatísticas	Idade
contagem	14830
média	41.950
Desvio padrão	17.367
mín.	6
25%	28
50%	40
75%	54
máx.	99

Podemos observar que a estatística descritiva dos anos das vítimas não apresenta mudanças perceptíveis. A média sempre girando em torno dos 42 anos, com o desvio padrão em torno de 17 e mediana girando nos 40 anos.

j) Suicídio por idade e sexo

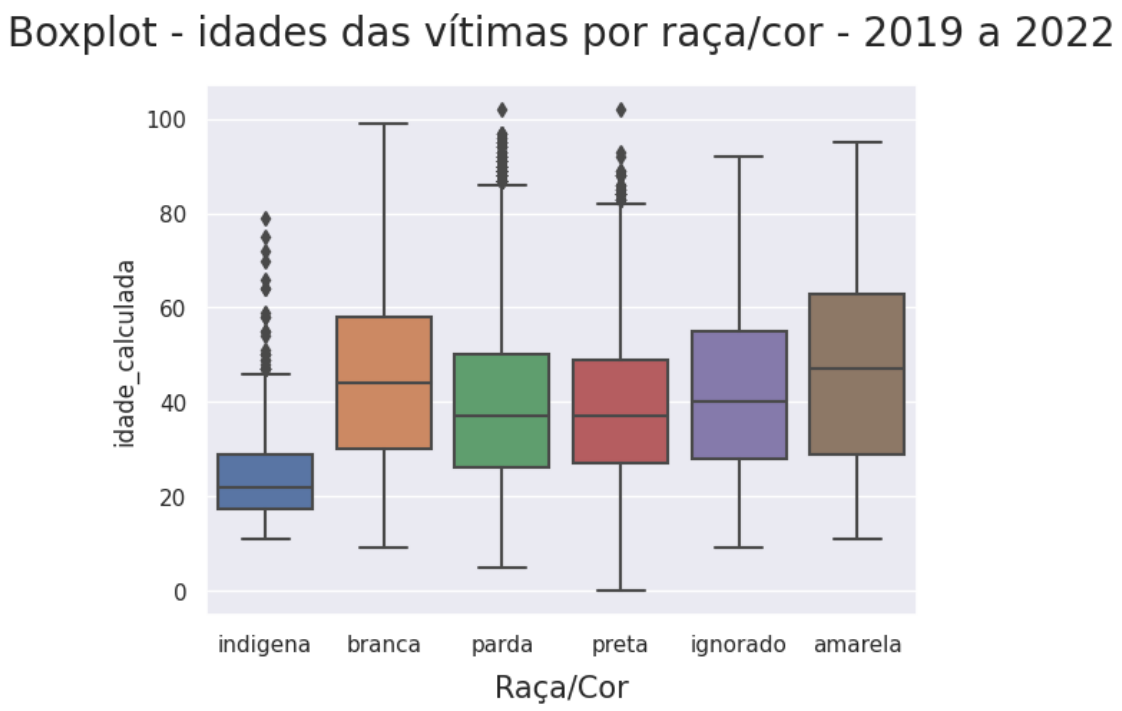
Tentou-se encontrar alguma relação entre idade e sexo, mas não foi possível identificar diferença a se destacar.



k) Suicídio por idade e raça/cor

Quanto à distribuição por idade separando as categorias de raça/cor, podemos identificar características a se destacar:

- Os indígenas cometem suicídio muito mais novos que os demais,
- Pardos e pretos têm a mesma característica de idade e
- Os amarelos, seguidos dos brancos, se suicidam mais velhos que os pretos e pardos.

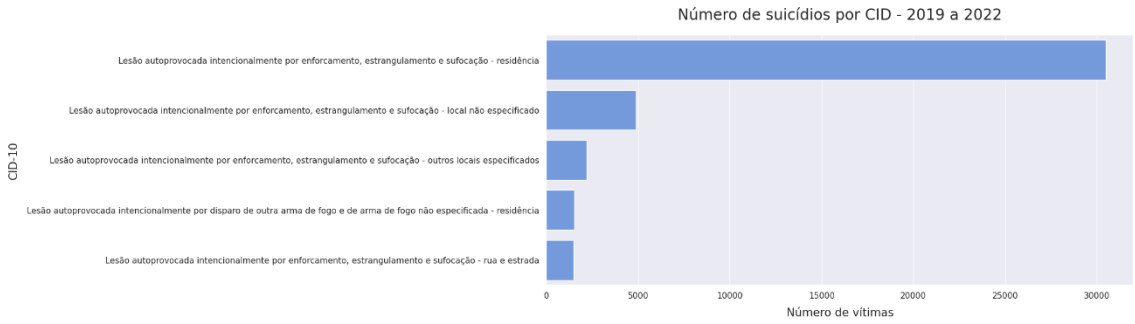


I) Suicídio pela causa básica - CID-10

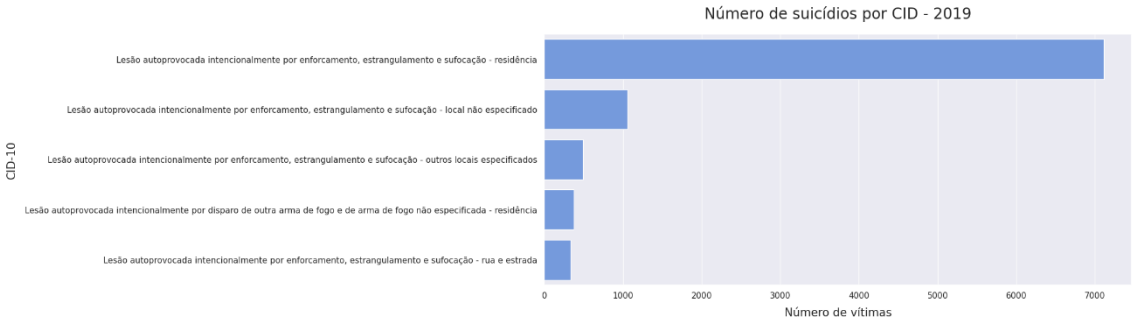
Em apertada síntese, CID-10 é a 10ª Revisão da Classificação Internacional de Doenças e problemas relacionados, trata-se de uma lista de classificação médica da Organização Mundial da Saúde (OMS) que contém os códigos para doenças, sinais e sintomas, achados anormais, queixas, circunstâncias sociais e causas externas de lesões ou doenças.

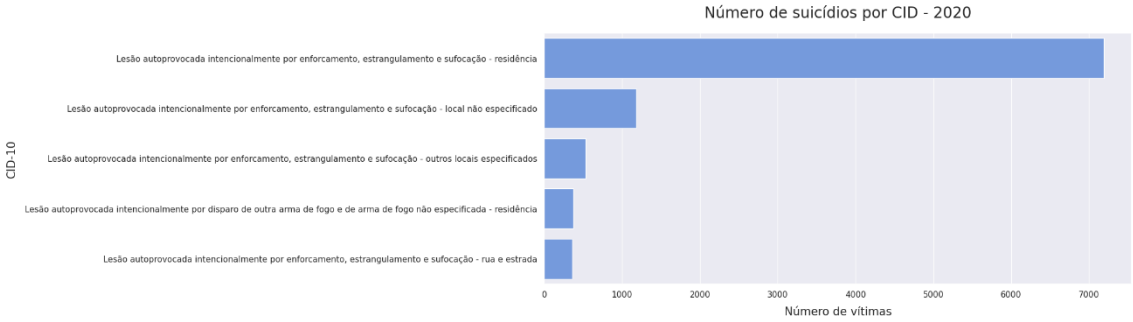
Neste ponto do estudo avisamos destacamos o **alerta de gatilho** e informamos que detalharemos como as pessoas escolhem se matar.

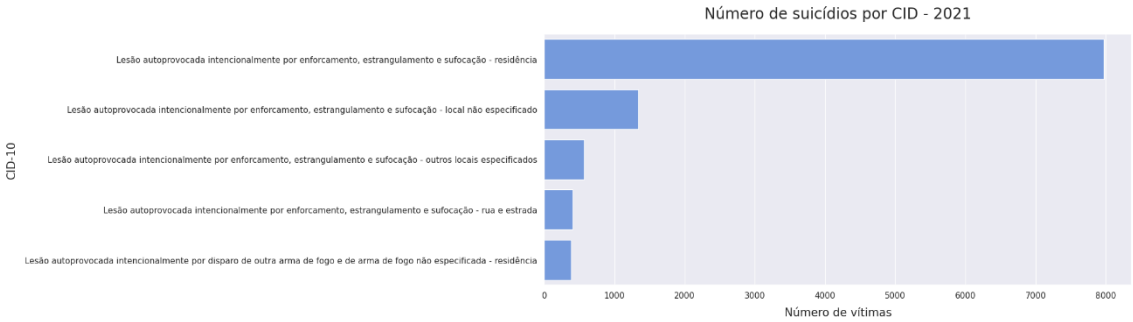
Conforme podemos observar, de 2019 a 2022 o principal método escolhido para se matar é o enforcamento, seguido de arma de fogo (2,7%) e precipitação de lugar elevado (1,61%). Destacando que a diferença pelo método de enforcamento é mais de 67% (3 primeiros da lista)

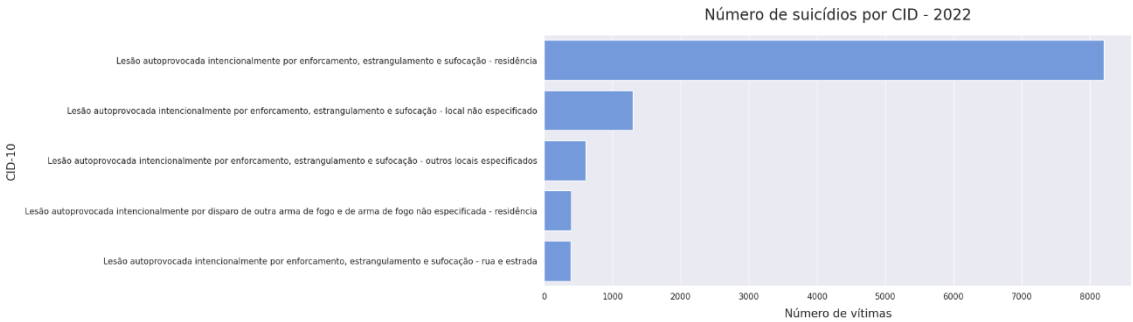


Não se encontrou diferença entre os anos quando analisados individualmente.









m) Suicídio pela causa básica - CID-10 e sexo

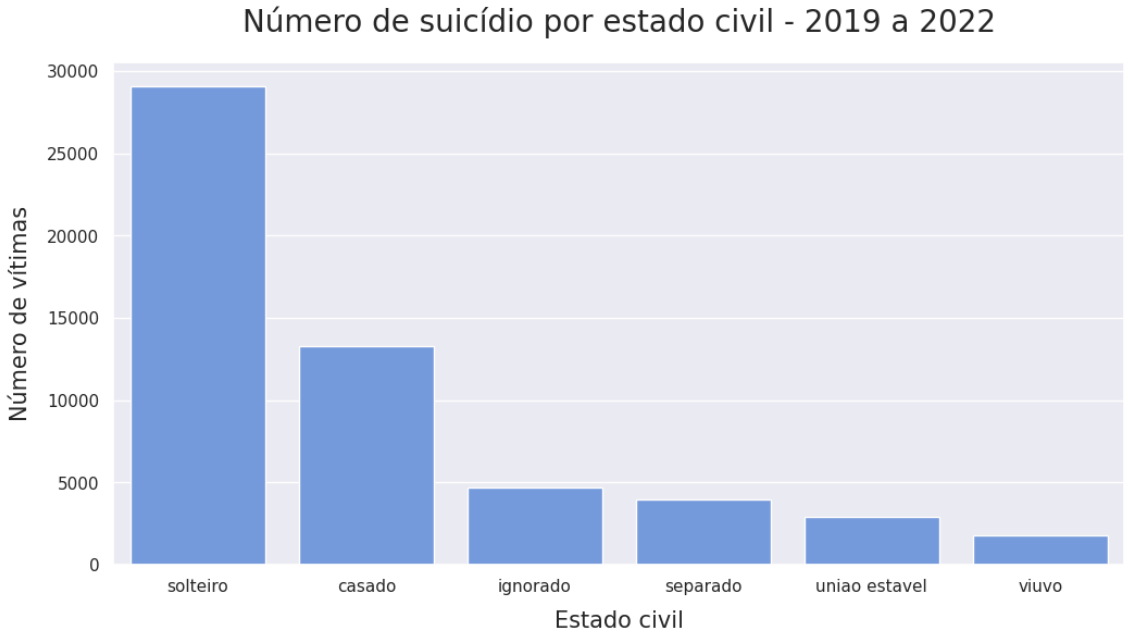
Pode-se destacar que não há diferença entre a escolha do principal método, mas há diferença entre escolha de métodos.

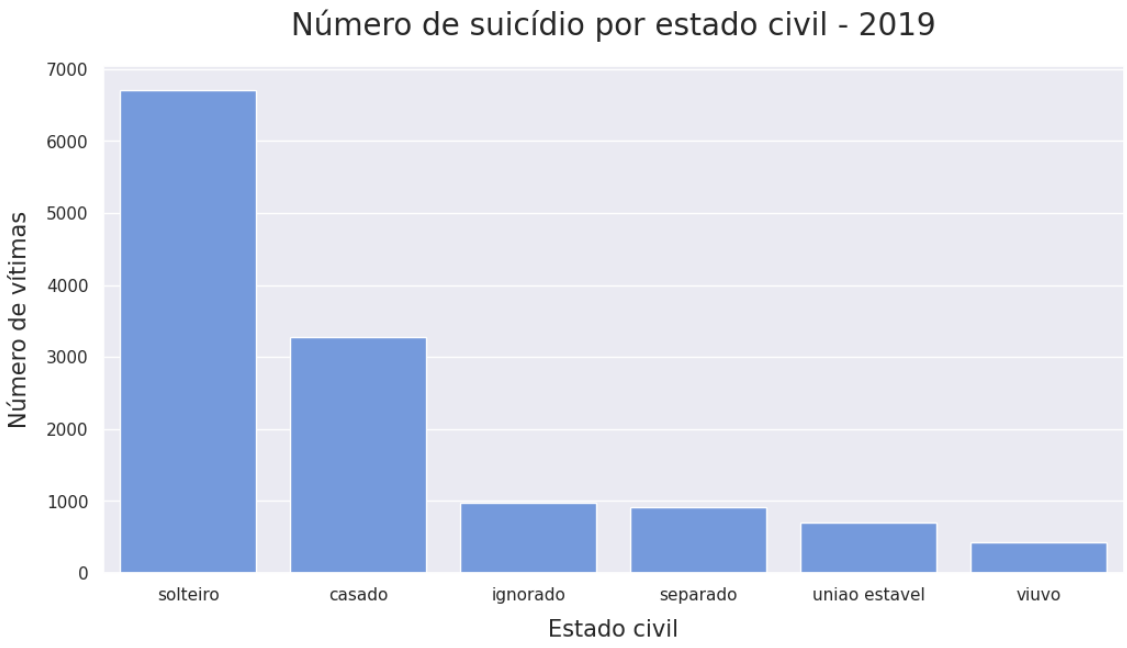
A Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e sufocação é a principal escolha de ambos os sexos, mas a

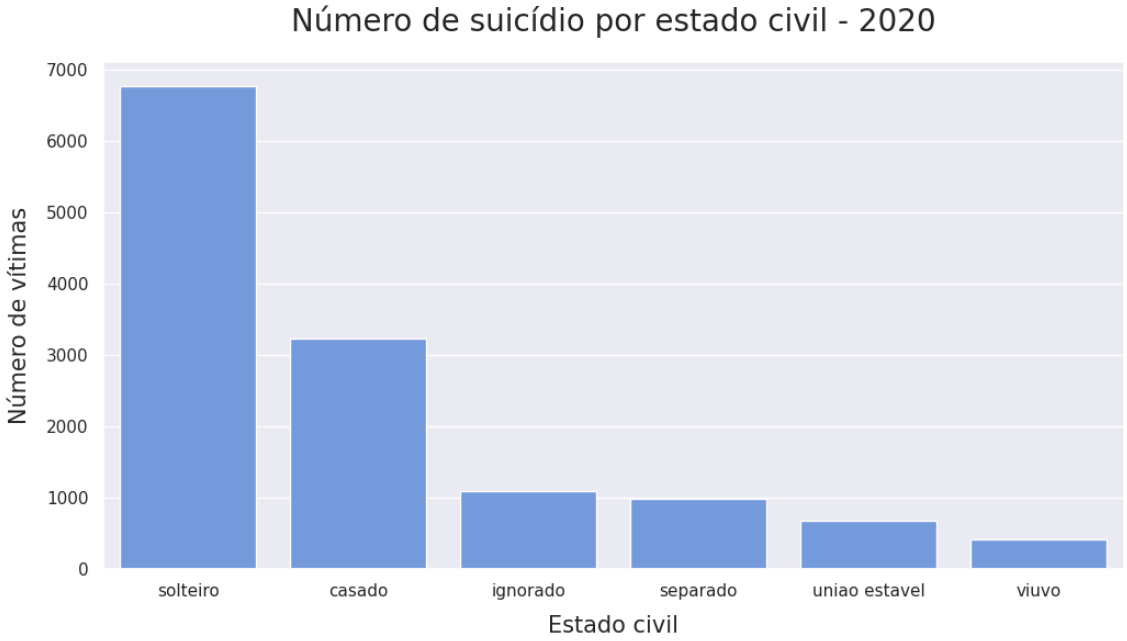
Autointoxicação por e exposição, intencional, a outras drogas, medicamentos e substâncias biológicas e às não especificadas é segunda forma mais prevalente entre as mulheres enquanto os homens preferem a Lesão autoprovocada intencionalmente por disparo de outra arma de fogo e de arma de fogo.

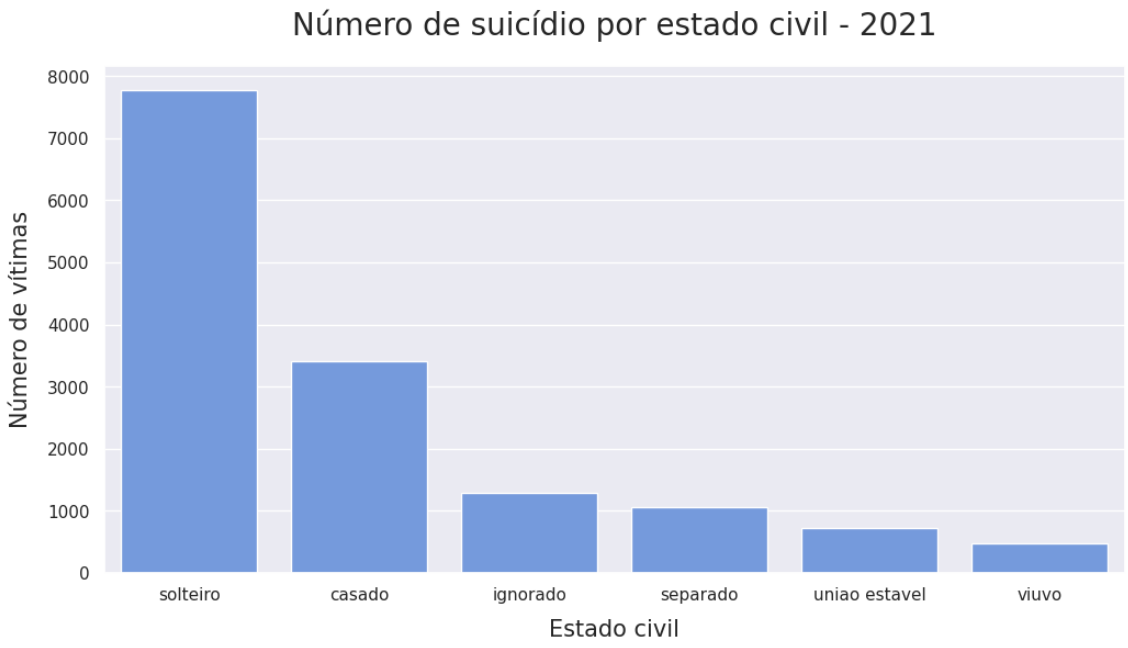
n) Suicídio por estado civil

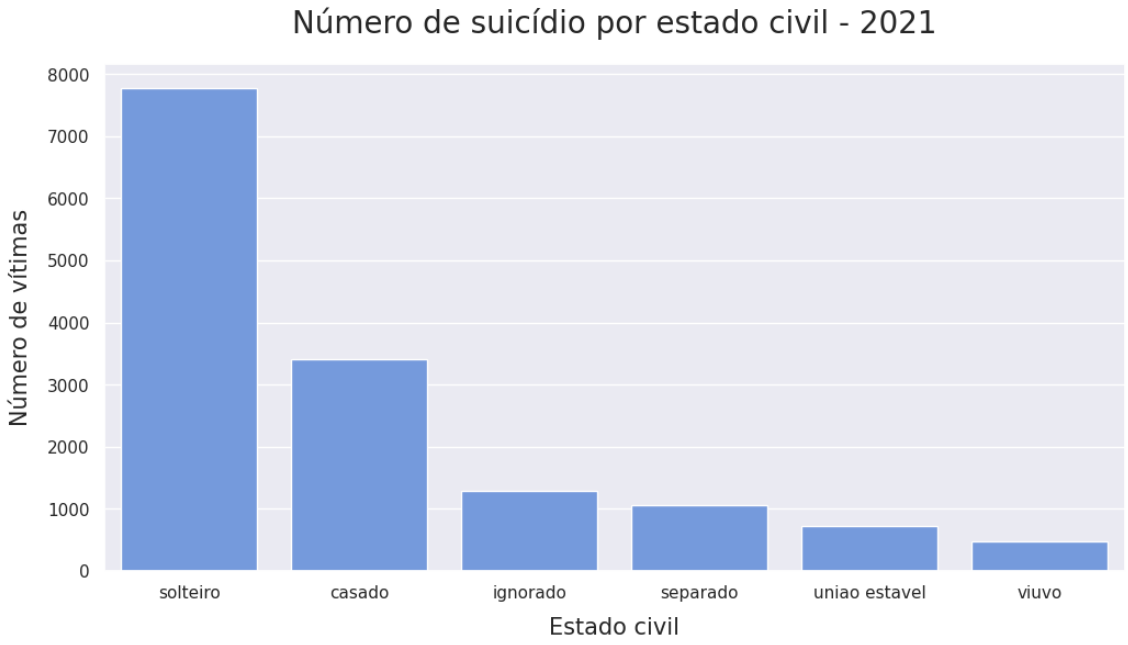
Nesta categoria comprova-se que mais de 50% dos suicídios são cometidos por solteiros e casados de 22 a 25%, e os demais, menos de 10%. Uma realidade que não muda pelos anos.




Número de suicídio por estado civil - 2019
Número de vítimas
7000
6000
5000
4000
3000
2000
1000
0
solteiro
casado
ignorado
separado
uniao estavel
viuvo
Estado civil


Número de suicídio por estado civil - 2020
Número de vítimas
7000
6000
5000
4000
3000
2000
1000
0
solteiro
casado
ignorado
separado
uniao estavel
viuvo
Estado civil





o) Suicídio por estado civil

Ao analisar se há diferença entre o estado civil em comparação a sexo, podemos destacar que a diferença entre homens e mulheres é menor quando o estado civil é separado.

SEXO	ESTCIV	CONTAGEM
masculino	**solteiro**	22882
	casado	10602
	ignorado	3788
	separado	2896
	união estável	2321
	viúvo	1144
ignorado	**ignorado**	9
	união estável	1
feminino	**solteiro**	6200
	casado	2672
	separado	1070
	ignorado	884
	viúvo	666
	união estável	592

XIV. Autoavaliação

A proposta de fazer um pipeline de dados com o uso de tecnologias em nuvem representou um desafio significativo, especialmente porque eu não tinha familiaridade prévia com as ferramentas disponíveis para essa finalidade. No entanto, com dedicação e pesquisa, acredito ter alcançado sucesso na construção de um pipeline funcional.

Minha escolha de utilizar a Google Cloud Platform como ambiente em nuvem me levou a explorar uma variedade de ferramentas, como o Google BigQuery, Cloud Storage, Data Fusion, Dataprep e Dataflow. Essas ferramentas foram essenciais para a extração, transformação e carga de dados em um formato adequado para a análise planejada. Embora eu já possuísse algum

conhecimento em Python e experiência em ambientes Linux, enfrentei desafios significativos ao lidar com o fluxo de trabalho e a utilização de algumas dessas ferramentas, com destaque para a Data Fusion.

Esta experiência se mostrou enriquecedora, proporcionando um aprendizado sólido e uma visão prática sobre a importância da engenharia de dados no contexto do trabalho de um cientista de dados. Além disso, ela destacou o imenso potencial das plataformas em nuvem para o processamento eficiente de grandes volumes de dados.

No final, acredito ter alcançado o objetivo que estabeleci no início deste projeto. Consegui analisar e descrever com sucesso as características das pessoas que cometem suicídio no país, examinando dados relacionados à escolaridade, sexo, unidade federativa, idade e até mesmo informações surpreendentes, como os horários e os métodos utilizados. Conseguindo, de maneira sintética, traçar que o perfil das pessoas que cometem suicídio no país é:

- Homem, branco ou pardo;
- Idade média de 42 anos;
- Baixa escolaridade (entre o fundamental completo e ensino médio incompleto);
- Solteiro;
- Se morador do interior, trabalhador agropecuário, se da cidade, pedreiro;
- Morador do estado de São Paulo, mas com atenção especial para o estado do Rio Grande do Sul e
- Escolhe retirar a própria vida por enforcamento, preferencialmente em casa.

Esta jornada me ensinou não apenas a importância da perseverança diante de desafios técnicos, mas também a relevância do campo da ciência de dados na abordagem de questões sociais complexas.

XV. Anexos

Receita de manipulação do Dataprep (Wrangler):

```
drop col: IDADEMAE, ESCMAE, ESCMAE2010, SERIESCMAE, OCUPMAE,
QTDFILVIVO, QTDFILMORT, GRAVIDEZ, SEMAGESTAC, GESTACAO, PARTO,
OBITOPARTO, PESO, TPMORTEOCO, ACIDTRAB action: Drop


drop col: OBITOGRAV, OBITOPUERP action: Drop


drop col: TIPOBITO, ORIGEM action: Drop


settype col: DTOBITO lockDataType: true type: String


settype col: HORAOBITO lockDataType: true type: String


columnbyexample columns: HORAOBITO value: CASE([ISNULL(HORAOBITO),
NULL(),ISMISSING([HORAOBITO]),'',MERGE([SUBSTRING(HORAOBITO, 0,
FINDNTH(HORAOBITO, /[0-9]/, 3, left, false)),':',SUBSTRING(HORAOBITO,
0, FINDNTH(HORAOBITO, /[0-9]/, 3, left, false))])]) exampleMappingId:
52265  as: 'HORAOBITO1'


drop col: HORAOBITO action: Drop.


rename type: manual mapping: [HORAOBITO1,'HORAOBITO']


columnbyexample columns: DTOBITO value:
CASE([ISNULL(DTOBITO),NULL(),ISMISSING([DTOBITO]),'',MERGE([SUBSTRING(
DTOBITO, 0, FINDNTH(DTOBITO, /[0-9]/, 3, left, false)),'-
',SUBSTRING(DTOBITO, LEN(DTOBITO) -  LEN(SUBSTITUTE(SUBSTRING(DTOBITO,
FINDNTH(DTOBITO, /[A-Za-z0-9]/, 7, right, false), LEN(DTOBITO)), /[A-
Za-z0-9]/, '', false, /^/)), FINDNTH(DTOBITO, /[A-Za-z0-9]/, 5, left,
false)),'-',SUBSTRING(DTOBITO, LEN(DTOBITO) -
LEN(SUBSTITUTE(SUBSTRING(DTOBITO, FINDNTH(DTOBITO, /[A-Za-z0-9]/, 4,
left, false), LEN(DTOBITO)), /[A-Za-z0-9]/, '', false, /^/)),
LEN(DTOBITO))])]) exampleMappingId: 52266  as: 'DTOBITO1'


dateformat col: DTOBITO1 type: slashdate


drop col: DTOBITO action: Drop


rename type: manual mapping: [DTOBITO1,'DTOBITO']


dateformat col: HORAOBITO type: custom format: 'HH:MM'
```

```
columnbyexample columns: DTNASC value:
CASE([ISNULL(DTNASC),NULL(),ISMISSING([DTNASC]),'',MERGE([SUBSTRING(DT
NASC, 0, FINDNTH(DTNASC, /[0-9]/, 3, left, false)),'-
',SUBSTRING(DTNASC, LEN(DTNASC) -  LEN(SUBSTITUTE(SUBSTRING(DTNASC,
FINDNTH(DTNASC, /[A-Za-z0-9]/, 7, right, false), LEN(DTNASC)), /[A-Za-
z0-9]/, '', false, /^/)), FINDNTH(DTNASC, /[A-Za-z0-9]/, 5, left,
false)),'-',SUBSTRING(DTNASC, LEN(DTNASC) -
LEN(SUBSTITUTE(SUBSTRING(DTNASC, FINDNTH(DTNASC, /[A-Za-z0-9]/, 4,
left, false), LEN(DTNASC)), /[A-Za-z0-9]/, '', false, /^/)),
LEN(DTNASC))])]) exampleMappingId: 52267  as: 'DTNASC1'


dateformat col: DTNASC1 type: slashdate


drop col: DTNASC action: Drop


rename type: manual mapping: [DTNASC1,'DTNASC']


columnbyexample columns: CODMUNOCOR value:
CASE([ISNULL(CODMUNOCOR),NULL(),ISMISSING([CODMUNOCOR]),'',SUBSTRING(C
ODMUNOCOR, 0, FINDNTH(CODMUNOCOR, /[0-9]/, 4, right, false))])
exampleMappingId: 52268  as: 'CODUFOCOR'


settype col: IDADE lockDataType: true type: Integer


columnbyexample columns: SEXO value:
CASE([ISNULL(SEXO),NULL(),ISMISSING([SEXO]),'',CASE([SEXO ==
'1','masculino',SEXO == '2','feminino',NULL()])]) exampleMappingId:
52273  as: 'SEXONOME'


drop col: SEXO action: Drop


rename type: manual mapping: [SEXONOME,'SEXO']


columnbyexample columns: RACACOR value:
CASE([ISNULL(RACACOR),NULL(),ISMISSING([RACACOR]),'',CASE([RACACOR ==
'1', 'branca', RACACOR == '2', 'preta', RACACOR == '3', 'amarela',
RACACOR == '4', 'parda', 'indigena'])]) exampleMappingId: 52274  as:
'RACACOR1'


drop col: RACACOR action: Drop


rename type: manual mapping: [RACACOR1,'RACACOR']


columnbyexample columns: ESTCIV value:
CASE([ISNULL(ESTCIV),NULL(),ISMISSING([ESTCIV]),'',CASE([ESTCIV ==
'2', 'casado', ESTCIV == '3', 'viuvo', ESTCIV == '4', 'separado',
ESTCIV == '5', 'uniao estavel', ESTCIV == '9', 'ignorado',
'solteiro'])]) exampleMappingId: 52275  as: 'ESTCIV1'
```

```
drop col: ESTCIV action: Drop


rename type: manual mapping: [ESTCIV1,'ESTCIV']


columnbyexample columns: ESC value:
CASE([ISNULL(ESC),NULL(),ISMISSING([ESC]),'',CASE([ESC == '2', 'de 1 a
3 anos', ESC == '3', 'de 4 a 7 anos', ESC == '4', 'de 8 a 11 anos',
ESC == '5', '12 anos e mais', ESC == '9', 'ignorado', 'nenhuma'])])
exampleMappingId: 52276  as: 'ESC1'


drop col: ESC action: Drop


rename type: manual mapping: [ESC1,'ESC']


columnbyexample columns: ESC2010 value:
CASE([ISNULL(ESC2010),NULL(),ISMISSING([ESC2010]),'',CASE([ESC2010 ==
'1', 'fundamental I (1ª a 4ª série)', ESC2010 == '2', 'fundamental II
(5ª a 8ª série)', ESC2010 == '3', 'médio (antigo 2° grau)', ESC2010 ==
'4', 'superior incompleto', ESC2010 == '5', 'superior completo',
ESC2010 == '9', 'ignorado', 'sem escolaridade'])]) exampleMappingId:
52277  as: 'ESC1'


drop col: ESC2010 action: Drop


rename type: manual mapping: [ESC1,'ESC2010']


columnbyexample columns: ASSISTMED value:
CASE([ISNULL(ASSISTMED),NULL(),ISMISSING([ASSISTMED]),'',CASE([ASSISTM
ED == '1','sim',ASSISTMED == '2','não','ignorado'])])
exampleMappingId: 52278  as: 'ASSISTMED1'


drop col: ASSISTMED action: Drop


rename type: manual mapping: [ASSISTMED1,'ASSISTMED']


columnbyexample columns: DTATESTADO value:
CASE([ISNULL(DTATESTADO),NULL(),ISMISSING([DTATESTADO]),'',MERGE([SUBS
TRING(DTATESTADO, 0, FINDNTH(DTATESTADO, /[0-9]/, 3, left, false)),'-
',SUBSTRING(DTATESTADO, LEN(DTATESTADO) -
LEN(SUBSTITUTE(SUBSTRING(DTATESTADO, FINDNTH(DTATESTADO, /[A-Za-z0-
9]/, 7, right, false), LEN(DTATESTADO)), /[A-Za-z0-9]/, '', false,
/^/)), FINDNTH(DTATESTADO, /[A-Za-z0-9]/, 5, left, false)),'-
',SUBSTRING(DTATESTADO, LEN(DTATESTADO) -
LEN(SUBSTITUTE(SUBSTRING(DTATESTADO, FINDNTH(DTATESTADO, /[A-Za-z0-
9]/, 4, left, false), LEN(DTATESTADO)), /[A-Za-z0-9]/, '', false,
/^/)), LEN(DTATESTADO))])]) exampleMappingId: 52279  as: 'DTATESTADO1'
```

```
drop col: DTATESTADO action: Drop


rename type: manual mapping: [DTATESTADO1,'DTATESTADO']


dateformat col: DTATESTADO type: slashdate


columnbyexample columns: CIRCOBITO value:
CASE([ISNULL(CIRCOBITO),NULL(),ISMISSING([CIRCOBITO]),'',CASE([CIRCOBI
TO == '2', 'suicídio', CIRCOBITO == '3', 'homicídio', CIRCOBITO ==
'4', 'outros', CIRCOBITO == '9', 'ignorado', 'acidente'])])
exampleMappingId: 52280  as: 'CIRCOBITO1'


drop col: CIRCOBITO action: Drop


rename type: manual mapping: [CIRCOBITO1,'CIRCOBITO']


columnbyexample columns: FONTE value:
CASE([ISNULL(FONTE),NULL(),ISMISSING([FONTE]),'',CASE([FONTE == '2',
'hospital', FONTE == '3', 'família', FONTE == '4', 'outra', FONTE ==
'9', 'ignorado', 'ocorrência policial'])]) exampleMappingId: 52281
as: 'FONTE1'


drop col: FONTE action: Drop


rename type: manual mapping: [FONTE1,'FONTE']


drop col: DTINVESTIG action: Drop


drop col: VERSAOSIST action: Drop


drop col: TPOBITOCOR action: Drop


drop col: NUDIASOBCO action: Drop


drop col: DIFDATA action: Drop


drop col: NUDIASOBIN action: Drop


drop col: DTCADINV action: Drop


drop col: DTCONINV action: Drop
```

```
drop col: ALTCAUSA, FONTESINF, DTCONCASO, MORTEPARTO, DTCADINF,
NUDIASINF, TPNIVELINV, TPRESGINFO action: Drop


drop col: FONTES action: Drop


drop col: STDOEPIDEM,STDONOVA action: Drop


drop col: CAUSAMAT action: Drop


drop col: ESCMAEAGR1 action: Drop


columnbyexample columns: ESCFALAGR1 value:
CASE([ISNULL(ESCFALAGR1),NULL(),ISMISSING([ESCFALAGR1]),'',CASE([ESCFA
LAGR1 == '01', 'fundamental I incompleto', ESCFALAGR1 == '02',
'fundamental I completo', ESCFALAGR1 == '03', 'fundamental II
incompleto', ESCFALAGR1 == '04', 'fundamental II completo', ESCFALAGR1
== '06', 'ensino médio completo', ESCFALAGR1 == '08', 'superior
completo', ESCFALAGR1 == '09', 'ignorado', ESCFALAGR1 == '10',
'fundamental I incompleto', ESCFALAGR1 == '11', 'fundamental II
incompleto', ESCFALAGR1 == '12', 'ensino médio incompleto', 'sem
escolaridade'])]) exampleMappingId: 52282  as: 'ESCFALAGR2'


drop col: ESCFALAGR1 action: Drop


rename type: manual mapping: [ESCFALAGR2,'ESCFALAGR1']


drop col: DTRECORIGA action: Drop


drop col: DTRECEBIM action: Drop


drop col: VERSAOSCB action: Drop


drop col: STCODIFICA action: Drop


drop col: DTCADASTRO,ATESTANTE,CODIFICADO,FONTEINV action: Drop


drop col: NUMEROLOTE action: Drop


drop col: ATESTADO action: Drop


drop col: CB_PRE action: Drop


drop col: LINHAA,LINHAB,LINHAC,LINHAD,LINHAII action: Drop
```

```
columnbyexample columns: CIRURGIA value:
CASE([ISNULL(CIRURGIA),NULL(),ISMISSING([CIRURGIA]),'',CASE([CIRURGIA
== '1','sim',CIRURGIA == '2','não','ignorado'])]) exampleMappingId:
52284  as: 'CIRURGIA1'


drop col: CIRURGIA action: Drop


rename type: manual mapping: [CIRURGIA1,'CIRURGIA']


drop col: CIRURGIA action: Drop


columnbyexample columns: NECROPSIA value:
CASE([ISNULL(NECROPSIA),NULL(),ISMISSING([NECROPSIA]),'',CASE([NECROPS
IA == '1','sim',NECROPSIA == '2','não','ignorado'])])
exampleMappingId: 52286  as: 'NECROPSIA1'


drop col: NECROPSIA action: Drop


rename type: manual mapping: [NECROPSIA1,'NECROPSIA']


drop col: EXAME action: Drop


drop col: ESTABDESCR action: Drop


drop col: CODESTAB action: Drop


columnbyexample columns: LOCOCOR value:
CASE([ISNULL(LOCOCOR),NULL(),ISMISSING([LOCOCOR]),'',CASE([LOCOCOR ==
'1', 'hospital', LOCOCOR == '2', 'outros estab. de saúde', LOCOCOR ==
'3', 'domicílio', LOCOCOR == '4', 'via pública', LOCOCOR == '5',
'outros', LOCOCOR == '9', 'ignorado', NULL()])]) exampleMappingId:
52288  as: 'LOCOCOR1'


drop col: LOCOCOR action: Drop


rename type: manual mapping: [LOCOCOR1,'LOCOCOR']


replacemissing col: HORAOBITO with: fill text: '' orderType: source


replacemissing col: SEXO with: custom text: 'ignorado'


replacemissing col: RACACOR with: custom text: 'ignorado'


replacemissing col: ESTCIV with: custom text: 'ignorado'
```

```
replacemissing col: ESC with: custom text: 'ignorado'


replacemissing col: ESC2010 with: custom text: 'ignorado'


replacemissing col: LOCOCOR with: custom text: 'ignorado'


replacemissing col: FONTE with: custom text: 'ignorado'


replacemissing col: ESCFALAGR1 with: custom text: 'ignorado'


replacemissing col: DTOBITO with: fill text: '' orderType: source


replacemissing col: DTNASC with: fill text: '' orderType: source


join  with: tabela_ocupacoes.csv col: OCUP = current_dataset.OCUP,
DTOBITO = current_dataset.DTOBITO, HORAOBITO =
current_dataset.HORAOBITO, NATURAL = current_dataset.NATURAL,
CODMUNNATU = current_dataset.CODMUNNATU, DTNASC =
current_dataset.DTNASC, IDADE = current_dataset.IDADE, SEXO =
current_dataset.SEXO, RACACOR = current_dataset.RACACOR, ESTCIV =
current_dataset.ESTCIV, ESC = current_dataset.ESC, ESC2010 =
current_dataset.ESC2010, SERIESCFAL = current_dataset.SERIESCFAL,
CODMUNRES = current_dataset.CODMUNRES, LOCOCOR =
current_dataset.LOCOCOR, CODMUNOCOR = current_dataset.CODMUNOCOR,
CODUFOCOR = current_dataset.CODUFOCOR, ASSISTMED =
current_dataset.ASSISTMED, NECROPSIA = current_dataset.NECROPSIA,
CAUSABAS = current_dataset.CAUSABAS, COMUNSVOIM =
current_dataset.COMUNSVOIM, DTATESTADO = current_dataset.DTATESTADO,
CIRCOBITO = current_dataset.CIRCOBITO, FONTE = current_dataset.FONTE,
TPPOS = current_dataset.TPPOS, CAUSABAS_O =
current_dataset.CAUSABAS_O, ESCFALAGR1 = current_dataset.ESCFALAGR1,
CONTADOR = current_dataset.CONTADOR, TITULO =
tabela_ocupacoes.csv.TITULO row: current_dataset.OCUP ==
tabela_ocupacoes.csv.CODIGO type: left


rename type: manual mapping: [TITULO,'TITULO_OCUPACAO']


replacemissing col: TITULO_OCUPACAO with: custom text: 'ignorado'


join  with: tabela_cid10 col: CAUSABAS = current_dataset.CAUSABAS,
OCUP = current_dataset.OCUP, DTOBITO = current_dataset.DTOBITO,
HORAOBITO = current_dataset.HORAOBITO, NATURAL =
current_dataset.NATURAL, CODMUNNATU = current_dataset.CODMUNNATU,
DTNASC = current_dataset.DTNASC, IDADE = current_dataset.IDADE, SEXO =
current_dataset.SEXO, RACACOR = current_dataset.RACACOR, ESTCIV =
current_dataset.ESTCIV, ESC = current_dataset.ESC, ESC2010 =
current_dataset.ESC2010, SERIESCFAL = current_dataset.SERIESCFAL,
CODMUNRES = current_dataset.CODMUNRES, LOCOCOR =
current_dataset.LOCOCOR, CODMUNOCOR = current_dataset.CODMUNOCOR,
```

```
CODUFOCOR = current_dataset.CODUFOCOR, ASSISTMED =
current_dataset.ASSISTMED, NECROPSIA = current_dataset.NECROPSIA,
COMUNSVOIM = current_dataset.COMUNSVOIM, DTATESTADO =
current_dataset.DTATESTADO, CIRCOBITO = current_dataset.CIRCOBITO,
FONTE = current_dataset.FONTE, TPPOS = current_dataset.TPPOS,
CAUSABAS_O = current_dataset.CAUSABAS_O, ESCFALAGR1 =
current_dataset.ESCFALAGR1, CONTADOR = current_dataset.CONTADOR,
TITULO_OCUPACAO = current_dataset.TITULO_OCUPACAO, DESCRICAO =
tabela_cid10.DESCRICAO row: current_dataset.CAUSABAS ==
tabela_cid10.SUBCAT type: left


join  with: ibge_codigo_UF.csv col: CODUFOCOR =
current_dataset.CODUFOCOR, CAUSABAS = current_dataset.CAUSABAS, OCUP =
current_dataset.OCUP, DTOBITO = current_dataset.DTOBITO, HORAOBITO =
current_dataset.HORAOBITO, NATURAL = current_dataset.NATURAL,
CODMUNNATU = current_dataset.CODMUNNATU, DTNASC =
current_dataset.DTNASC, IDADE = current_dataset.IDADE, SEXO =
current_dataset.SEXO, RACACOR = current_dataset.RACACOR, ESTCIV =
current_dataset.ESTCIV, ESC = current_dataset.ESC, ESC2010 =
current_dataset.ESC2010, SERIESCFAL = current_dataset.SERIESCFAL,
CODMUNRES = current_dataset.CODMUNRES, LOCOCOR =
current_dataset.LOCOCOR, CODMUNOCOR = current_dataset.CODMUNOCOR,
ASSISTMED = current_dataset.ASSISTMED, NECROPSIA =
current_dataset.NECROPSIA, COMUNSVOIM = current_dataset.COMUNSVOIM,
DTATESTADO = current_dataset.DTATESTADO, CIRCOBITO =
current_dataset.CIRCOBITO, FONTE = current_dataset.FONTE, TPPOS =
current_dataset.TPPOS, CAUSABAS_O = current_dataset.CAUSABAS_O,
ESCFALAGR1 = current_dataset.ESCFALAGR1, CONTADOR =
current_dataset.CONTADOR, TITULO_OCUPACAO =
current_dataset.TITULO_OCUPACAO, DESCRICAO =
current_dataset.DESCRICAO, Nome_UF = ibge_codigo_UF.csv.Nome_UF row:
current_dataset.CODUFOCOR == ibge_codigo_UF.csv.UF type: left


rename type: manual mapping: [DESCRICAO,'CID10']


rename type: manual mapping: [Nome_UF,'UFOCOR']
```

MVP_03_Análise_dos_dados_Com_modificações

October 1, 2023

1 Trabalho de análise dos óbitos causados por suicídio do anos de 2019 à 2022, registrados no DATASUS.

1.0.1 Setembro amarelo

Trabalho desenvolvido por: *Eduardo Ferreira Gonçalves*

Observação: Notebook modificado para reprodutibilidade sem a necessidade de acesso ao Google cloud ou BigQuery

---

1.1 Importação das bibliotecas e configuração para uso dos dados disponibilizados no BigQuery

[ ]:
```
# Somente necessário quando for processar os dados provenientes diretamente do␣
↪BigQuery
#from google.colab import auth
#from google.cloud import bigquery
#from google.colab import data_table

#project = 'mvp-puc-399114' # Project ID inserted based on the query results␣
↪selected to explore
#location = 'US' # Location inserted based on the query results selected to␣
↪explore
#client = bigquery.Client(project=project, location=location)
#data_table.enable_dataframe_formatter()
#auth.authenticate_user()
```

[ ]:
```
# Import de manipulação
import pandas as pd
import numpy as np

# import de visualização
import matplotlib.pyplot as plt
import seaborn as sns

# import complementar para desabilitar warnings
```

```python
import warnings

# configurações
# configurar o pandas para mostrar todas as colunas/atributos do dataframe
pd.set_option('display.max_columns', None)

#ocultar os warnings
warnings.filterwarnings("ignore")

# Deixar os gráficos do pyplot com o mesmo tema do seaborn
plt.style.use("seaborn-v0_8-darkgrid")

# Colocar o tema padão do seaborn
sns.set_theme(style='darkgrid')
```

1.2 Funções personalizadas

[ ]:
```python
# Transformado em dataframe e ordenado para melhor apresentação no gráfico
def gerar_dataframe_categorico(atributo_categorico:pd.Series, nome_categoria:
 ↪str, nome_coluna_somada:str):
    """
        Função para gerar um dataframe com a coluna categorica fornecida com a␣
 ↪contagem de linhas de cada categoria
    """
    dataframe_categorico = atributo_categorico.value_counts()
    dataframe_categorico = pd.DataFrame(dataframe_categorico).reset_index()
    dataframe_categorico.columns = [nome_categoria, nome_coluna_somada]
    return dataframe_categorico
```

[ ]:
```python
def grafico_de_colunas(x:pd.Series, y:pd.Series, dados:pd.DataFrame,␣
 ↪nome_grafico:str, rotulo_x:str, rotulo_y:str):
    """
        Função para facilitar e padronizar os gráficos utilizando o seaborn.
        função sns.barplot
    """
    plt.figure(figsize = (12, 6))
    plot = sns.barplot(x= x, y= y, data= dados, color='#6495ED')
    plt.title(nome_grafico, fontsize=20, pad=20)
    plt.xlabel(rotulo_x, fontsize=15, labelpad=10)
    plt.ylabel(rotulo_y, fontsize=15, labelpad=10)
    plt.show()
```

[ ]:
```python
def grafico_de_barras(x:pd.Series, y:pd.Series, dados:pd.DataFrame,␣
 ↪nome_grafico:str, rotulo_x:str, rotulo_y:str):
    """
        Função para facilitar e padronizar os gráficos utilizando o seaborn.
```

```
        função sns.barplot com orientação horizontal
    """
    plt.figure(figsize = (14, 6))
    plot = sns.barplot(x= x, y= y, data= dados, color='#6495ED', orient='h')
    plt.title(nome_grafico, fontsize=20, pad=20)
    plt.xlabel(rotulo_x, fontsize=15, labelpad=10)
    plt.ylabel(rotulo_y, fontsize=15, labelpad=10)
    plt.show()
```

[ ]:
```
def grafico_de_linha_temporal(x:pd.Series, y:pd.Series, dados:pd.DataFrame,␣
↪nome_grafico:str, rotulo_x:str, rotulo_y:str):
    """
        Função para facilitar e padronizar os gráficos utilizando o seaborn.
        função sns.lineplot
    """
    plt.figure(figsize = (16, 6))
    sns.lineplot(y=y, x=x, data=dados)
    plt.title(nome_grafico, fontsize=20, pad=20)
    plt.xlabel(rotulo_x, fontsize=15, labelpad=10)
    plt.ylabel(rotulo_y, fontsize=15, labelpad=10)
    plt.show()
```

[ ]:
```
def grafico_de_caixa(x:pd.Series, y:pd.Series, dados:pd.DataFrame, nome_grafico:
↪str, rotulo_x:str):
    """
        Função para facilitar e padronizar os gráficos utilizando o seaborn.
        função sns.boxplot
    """
    sns.boxplot(y=y, x=x, data=dados)
    plt.title(nome_grafico,fontsize=20, pad=20)
    plt.xlabel(rotulo_x, fontsize=15, labelpad=10)
    plt.show()
```

1.3 Conjunto de resultados carregado do trabalho do BigQuery como um DataFrame

Os resultados da consulta são referenciados pelo ID do job executado no BigQuery e pela consulta não precisa ser executado novamente para explorar os resultados. O método `to_dataframe` baixa os resultados para um DataFrame do Pandas usando a API BigQuery Storage.

[ ]:
```
# método de consumo via Bigquery
#job = client.get_job('bquxjob_169ce87c_18adbd5612a')
#dados = job.to_dataframe()
#dados
```

[ ]:
```
```

```python
# método através do github
dados = pd.read_csv('https://raw.githubusercontent.com/dudu1626/MVP-Sprint-03/
↪main/suicidio_completo.csv')
dados
```

[ ]:

```
       CODUFOCOR CAUSABAS      OCUP     DTOBITO HORAOBITO  NATURAL  \
0             33     X709  715210.0  2021-10-04  23:01:00    825.0
1             33     X700  519205.0  2022-03-21  06:01:00    833.0
2             22     X700  715615.0  2019-11-18  11:01:00    822.0
3             35     X700  715210.0  2020-04-08  17:01:00    835.0
4             35     X704  782315.0  2020-04-11  04:01:00    835.0
…           …      …        …          …        …       …
55722         32     X706  999994.0  2020-02-01  20:01:00    833.0
55723         32     X700  715210.0  2020-09-10  09:01:00    832.0
55724         26     X709  621005.0  2022-05-21  21:01:00    826.0
55725         26     X700  999992.0  2020-09-10  15:01:00    826.0
55726         23     X648  622005.0  2022-09-08  12:01:00    823.0

       CODMUNNATU      DTNASC  IDADE       SEXO   RACACOR    ESTCIV  \
0        251530.0  1968-04-22    453  masculino     parda  solteiro
1        330590.0  1969-01-26    453  masculino    branca  solteiro
2        221100.0  1990-07-28    429  masculino  ignorado  ignorado
3        351290.0  1966-08-04    453  masculino    branca  ignorado
4        352320.0  1952-11-02    467  masculino     parda    casado
…           …         …    …         …       …       …
55722    330045.0  1978-11-02    441  masculino     preta  solteiro
55723    320530.0  1990-09-25    429  masculino     parda  solteiro
55724    261330.0  2004-10-15    417  masculino     parda  solteiro
55725    260520.0  1989-08-08    431   feminino     parda  solteiro
55726    230655.0  1983-07-23    439  masculino     parda  solteiro

                 ESC                          ESC2010  SERIESCFAL  CODMUNRES  \
0           ignorado                         ignorado         NaN     330240
1           ignorado                         ignorado         NaN     330240
2           ignorado                         ignorado         NaN     220160
3           ignorado                         ignorado         NaN     353280
4           ignorado                         ignorado         NaN     353280
…              …                                 …         …         …
55722  de 4 a 7 anos  fundamental II (5ª a 8ª série)         NaN     320510
55723  de 4 a 7 anos  fundamental II (5ª a 8ª série)         NaN     320510
55724  de 4 a 7 anos  fundamental II (5ª a 8ª série)         NaN     260350
55725  de 4 a 7 anos  fundamental II (5ª a 8ª série)         NaN     261630
55726  de 4 a 7 anos  fundamental II (5ª a 8ª série)         NaN     230655

           LOCOCOR  CODMUNOCOR ASSISTMED NECROPSIA  COMUNSVOIM  DTATESTADO  \
0           outros      330240       não       sim    330240.0  2021-10-05
1           outros      330240       não       sim    330240.0  2022-03-21
```

2	outros	220160	não	sim	221100.0	2019-11-20
3	domicílio	353280	não	sim	354980.0	2020-04-09
4	via pública	353280	não	sim	354980.0	2020-04-11
…	…	…	…	…	…	…
55722	outros	320510	não	sim	320530.0	2020-02-01
55723	domicílio	320510	não	sim	320530.0	2020-09-10
55724	outros	260350	NaN	NaN	260410.0	2022-05-21
55725	domicílio	261630	não	não	261160.0	2020-09-10
55726	outros	230655	sim	sim	230640.0	2022-09-09

	CIRCOBITO	FONTE	TPPOS	CAUSABAS_O	\
0	suicídio	ocorrência policial	False	X709	
1	suicídio	ocorrência policial	NaN	X709	
2	suicídio	família	NaN	X700	
3	suicídio	ocorrência policial	False	X700	
4	suicídio	ocorrência policial	False	X704	
…	…	…	…	…	
55722	suicídio	ocorrência policial	NaN	Y206	
55723	suicídio	ocorrência policial	False	X700	
55724	suicídio	ocorrência policial	NaN	X709	
55725	suicídio	ocorrência policial	NaN	Y340	
55726	suicídio	ocorrência policial	NaN	X648	

	ESCFALAGR1	CONTADOR	TITULO_OCUPACAO	\
0	ignorado	1769230	Pedreiro	
1	ignorado	95178	Catador de material reciclável	
2	ignorado	752341	Eletricista de instalações	
3	ignorado	357856	Pedreiro	
4	ignorado	357865	Motorista de táxi	
…	…	…	…	
55722	fundamental II incompleto	307141	ignorado	
55723	fundamental II incompleto	1135921	Pedreiro	
55724	fundamental II incompleto	509558	Trabalhador agropecuário em geral	
55725	fundamental II incompleto	1107478	ignorado	
55726	fundamental II incompleto	1098900	Caseiro (agricultura)	

	CID10	UFOCOR	ANO
0	Lesão autoprovocada intencionalmente por enfor…	Rio de Janeiro	2021
1	Lesão autoprovocada intencionalmente por enfor…	Rio de Janeiro	2022
2	Lesão autoprovocada intencionalmente por enfor…	Piauí	2019
3	Lesão autoprovocada intencionalmente por enfor…	São Paulo	2020
4	Lesão autoprovocada intencionalmente por enfor…	São Paulo	2020
…	…	…	…
55722	Lesão autoprovocada intencionalmente por enfor…	Espírito Santo	2020
55723	Lesão autoprovocada intencionalmente por enfor…	Espírito Santo	2020
55724	Lesão autoprovocada intencionalmente por enfor…	Pernambuco	2022
55725	Lesão autoprovocada intencionalmente por enfor…	Pernambuco	2020

```
55726  Auto-intoxicação por e exposição, intencional,…          Ceará  2022

[55727 rows x 32 columns]
```

1.4 Qualidade dos dados

1.4.1 Dimensões da tabela

[ ]:
```
dados.shape
```

[ ]:
```
(55727, 32)
```

1.4.2 Observar quais variáveis estão disponíveis, tipo dos dados e quantidade de nulos.

[ ]:
```
dados.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55727 entries, 0 to 55726
Data columns (total 32 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   CODUFOCOR        55727 non-null  int64
 1   CAUSABAS         55727 non-null  object
 2   OCUP             46272 non-null  float64
 3   DTOBITO          55727 non-null  object
 4   HORAOBITO        55727 non-null  object
 5   NATURAL          54176 non-null  float64
 6   CODMUNNATU       53785 non-null  float64
 7   DTNASC           55727 non-null  object
 8   IDADE            55727 non-null  int64
 9   SEXO             55727 non-null  object
 10  RACACOR          55727 non-null  object
 11  ESTCIV           55727 non-null  object
 12  ESC              55727 non-null  object
 13  ESC2010          55727 non-null  object
 14  SERIESCFAL       18530 non-null  float64
 15  CODMUNRES        55727 non-null  int64
 16  LOCOCOR          55727 non-null  object
 17  CODMUNOCOR       55727 non-null  int64
 18  ASSISTMED        46062 non-null  object
 19  NECROPSIA        48819 non-null  object
 20  COMUNSVOIM       50205 non-null  float64
 21  DTATESTADO       55566 non-null  object
 22  CIRCOBITO        55727 non-null  object
 23  FONTE            55727 non-null  object
 24  TPPOS            23807 non-null  object
 25  CAUSABAS_O       55668 non-null  object
```

```
 26  ESCFALAGR1       55727 non-null  object
 27  CONTADOR         55727 non-null  int64
 28  TITULO_OCUPACAO  55727 non-null  object
 29  CID10            55727 non-null  object
 30  UFOCOR           55727 non-null  object
 31  ANO              55727 non-null  int64
dtypes: float64(5), int64(6), object(21)
memory usage: 13.6+ MB
```

1.5 Ordenamento da tabela por data

[ ]:
```
dados = dados.sort_values(by='DTOBITO').reset_index(drop=True)
dados
```

[ ]:
```
       CODUFOCOR CAUSABAS      OCUP     DTOBITO HORAOBITO  NATURAL  \
0             13     X700  612005.0  2019-01-01  18:01:00    813.0
1             51     X700  999991.0  2019-01-01  05:01:00    851.0
2             28     X700  999991.0  2019-01-01  06:01:00    829.0
3             13     X700  999994.0  2019-01-01  02:01:00    813.0
4             16     X702  999991.0  2019-01-01  20:01:00    815.0
…          …      …        …         …        …      …
55722         52     X840  141410.0  2022-12-31  03:01:00    831.0
55723         29     X700  621005.0  2022-12-31  13:01:00    829.0
55724         42     X700  782510.0  2022-12-31  23:01:00    841.0
55725         43     X700       NaN  2022-12-31  18:01:00    843.0
55726         41     X700  513405.0  2022-12-31  04:01:00    841.0

       CODMUNNATU      DTNASC  IDADE       SEXO   RACACOR         ESTCIV  \
0        130006.0  1994-09-07    424  masculino  indigena       solteiro
1        510395.0  2004-09-11    414   feminino    branca       solteiro
2        292400.0  2000-07-03    418  masculino     parda       solteiro
3        130290.0  1985-11-28    433  masculino     parda       solteiro
4        150050.0  1997-01-24    421   feminino     parda       solteiro
…           …         …    …         …       …            …
55722    314800.0  1989-06-16    433  masculino    branca       ignorado
55723    290380.0  1997-03-05    425  masculino     parda       solteiro
55724    412230.0  1978-05-12    444  masculino    branca  uniao estavel
55725    430435.0  1989-08-07    433  masculino    branca       solteiro
55726    411960.0  1992-05-16    430  masculino    branca       solteiro

                  ESC                        ESC2010  SERIESCFAL  CODMUNRES  \
0      de 8 a 11 anos          médio (antigo 2º grau)         NaN     130006
1      de 8 a 11 anos          médio (antigo 2º grau)         NaN     510395
2       de 4 a 7 anos  fundamental II (5ª a 8ª série)         5.0     280450
3             nenhuma                sem escolaridade         NaN     130290
4      de 8 a 11 anos          médio (antigo 2º grau)         NaN     150050
…               …                                …          …          …
```

```
55722        ignorado                        ignorado         NaN     530010
55723   de 4 a 7 anos  fundamental II (5ª a 8ª série)         NaN     290380
55724        ignorado                        ignorado         NaN     421010
55725        ignorado                        ignorado         NaN     431100
55726  de 8 a 11 anos          médio (antigo 2º grau)         NaN     411960

         LOCOCOR  CODMUNOCOR ASSISTMED NECROPSIA  COMUNSVOIM  DTATESTADO  \
0      domicílio      130006       não       não         NaN  2019-01-03
1      domicílio      510395       NaN       NaN    510250.0  2019-01-01
2      domicílio      280450       não       sim    280030.0  2019-01-01
3      domicílio      130290       sim       não         NaN  2019-01-01
4       hospital      160030       não       sim    160030.0  2019-01-02
…          …         …       …       …         …         …
55722     outros      520450       não       sim    521380.0  2022-12-31
55723  domicílio      290380       não       sim    291470.0  2023-01-02
55724  domicílio      421010       sim       não    420380.0  2023-01-01
55725  domicílio      431100       não       sim    431440.0  2023-01-01
55726  domicílio      411960       não       sim    411150.0  2022-12-31

      CIRCOBITO                FONTE  TPPOS CAUSABAS_O  \
0      suicídio              família    NaN       X700
1      suicídio  ocorrência policial  False       X700
2      suicídio  ocorrência policial  False       X700
3      suicídio              família    NaN       X700
4      suicídio  ocorrência policial    NaN       X702
…         …                  …    …        …
55722  suicídio  ocorrência policial  False       X840
55723  suicídio  ocorrência policial  False       X700
55724  suicídio                outra  False       X700
55725  suicídio  ocorrência policial  False       X700
55726  suicídio                outra  False       X700

                      ESCFALAGR1  CONTADOR  \
0        ensino médio incompleto    510709
1        ensino médio incompleto    185795
2      fundamental II incompleto    733808
3               sem escolaridade    137583
4        ensino médio incompleto    278318
…                            …       …
55722                   ignorado   1539860
55723  fundamental II incompleto   1539640
55724                   ignorado   1539615
55725                   ignorado   1540657
55726    ensino médio incompleto   1539112

                                        TITULO_OCUPACAO  \
0                          Produtor agrícola polivalente
```

```
1                                                ignorado
2                                                ignorado
3                                                ignorado
4                                                ignorado
…                                                   …
55722                                Comerciante varejista
55723                   Trabalhador agropecuário em geral
55724  Motorista de caminhão (rotas regionais e inter…
55725                                            ignorado
55726                                              Garçom

                                                   CID10             UFOCOR  \
0      Lesão autoprovocada intencionalmente por enfor…           Amazonas
1      Lesão autoprovocada intencionalmente por enfor…        Mato Grosso
2      Lesão autoprovocada intencionalmente por enfor…            Sergipe
3      Lesão autoprovocada intencionalmente por enfor…           Amazonas
4      Lesão autoprovocada intencionalmente por enfor…              Amapá
…                                                   …                …
55722  Lesão autoprovocada intencionalmente por meios…              Goiás
55723  Lesão autoprovocada intencionalmente por enfor…              Bahia
55724  Lesão autoprovocada intencionalmente por enfor…     Santa Catarina
55725  Lesão autoprovocada intencionalmente por enfor…  Rio Grande do Sul
55726  Lesão autoprovocada intencionalmente por enfor…             Paraná

        ANO
0      2019
1      2019
2      2019
3      2019
4      2019
…     …
55722  2022
55723  2022
55724  2022
55725  2022
55726  2022

[55727 rows x 32 columns]
```

1.6 Separação dos dados por ano para análises separadas

[ ]:
```
dados_2019 = dados[dados.ANO == 2019]
dados_2020 = dados[dados.ANO == 2020]
dados_2021 = dados[dados.ANO == 2021]
dados_2022 = dados[dados.ANO == 2022]
```

2 Análise do perfil das pessoas que cometeram o suicídio

2.1 Análise dos suicídios pelo tempo

[ ]:
```
suicidios_por_ano = dados[['DTOBITO', 'ANO']].groupby('ANO').count()
#usando DTOBITO por ser um dado obrigatório na tabela
suicidios_por_ano.columns = ['NUM_SUICIDIOS']
suicidios_por_ano
```

[ ]:

	NUM_SUICIDIOS
ANO	
2019	13009
2020	13166
2021	14722
2022	14830

Podemos perceber um aumento no número de suicídios conforme os anos passam. Seria razoável se considerar a pandemia, mas o aumento expressivo se deu entre 2020 e 2021, mas em 2022 aumentou ainda mais.

[ ]:
```
# valores de óbitos de causas externas obtido em uma consulta no BigQuery
causas_externas = pd.DataFrame([
    (2019,142800),
    (2020,146038),
    (2021,149322),
    (2022,152090)],
    columns=['ANO', 'NUM_OBITOS'])

suicidios_por_ano = suicidios_por_ano.join(causas_externas.set_index('ANO'))
suicidios_por_ano['PERC_SUICIDIOS'] = (suicidios_por_ano.NUM_SUICIDIOS /␣
 ↪suicidios_por_ano.NUM_OBITOS) * 100
suicidios_por_ano
```

[ ]:

	NUM_SUICIDIOS	NUM_OBITOS	PERC_SUICIDIOS
ANO			
2019	13009	142800	9.109944
2020	13166	146038	9.015462
2021	14722	149322	9.859230
2022	14830	152090	9.750805

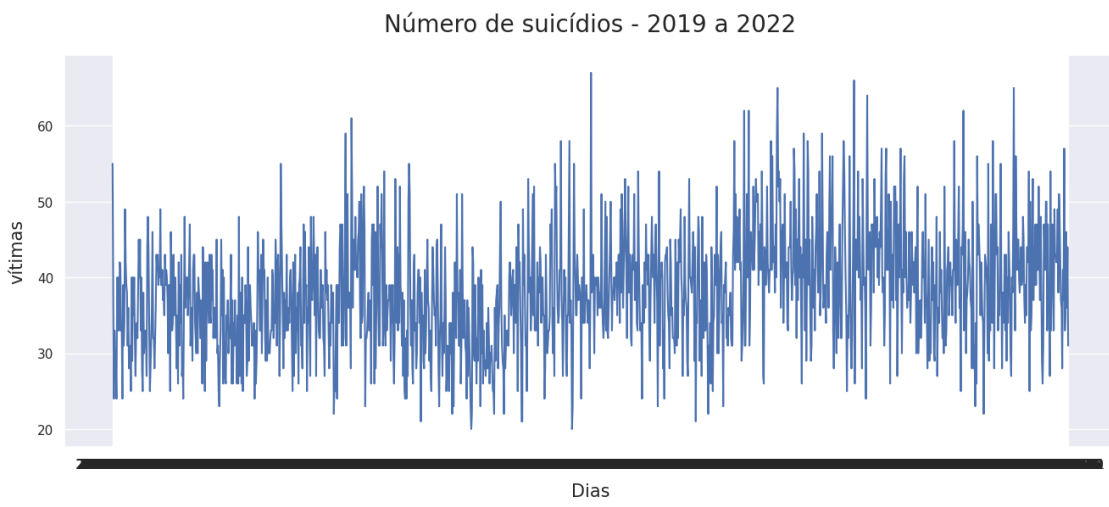
Considerando o número relativo do número de suicídios em relação ao número de mortes de causas externas totais, podemos perceber que houve um aumento de 9% de 2020 (primeiro ano da pandemia) para quase 10% em 2021 e 2022.

2.1.1 Série temporal

[ ]:
```
suicidio_por_data = gerar_dataframe_categorico(dados.DTOBITO, 'dia', 'contagem')
suicidio_por_data = suicidio_por_data.sort_values('dia')
suicidio_por_data
```

[ ]:
```
             dia  contagem
39    2019-01-01        55
725   2019-01-02        38
1432  2019-01-03        24
1062  2019-01-04        33
1280  2019-01-05        29
…            …       …
876   2022-12-27        36
213   2022-12-28        46
880   2022-12-29        36
296   2022-12-30        44
1210  2022-12-31        31

[1461 rows x 2 columns]
```

[ ]:
```
grafico_de_linha_temporal(x='dia', y='contagem', dados=suicidio_por_data,␣
↪nome_grafico="Número de suicídios - 2019 a 2022",
                          rotulo_x="Dias", rotulo_y="vítimas")
```



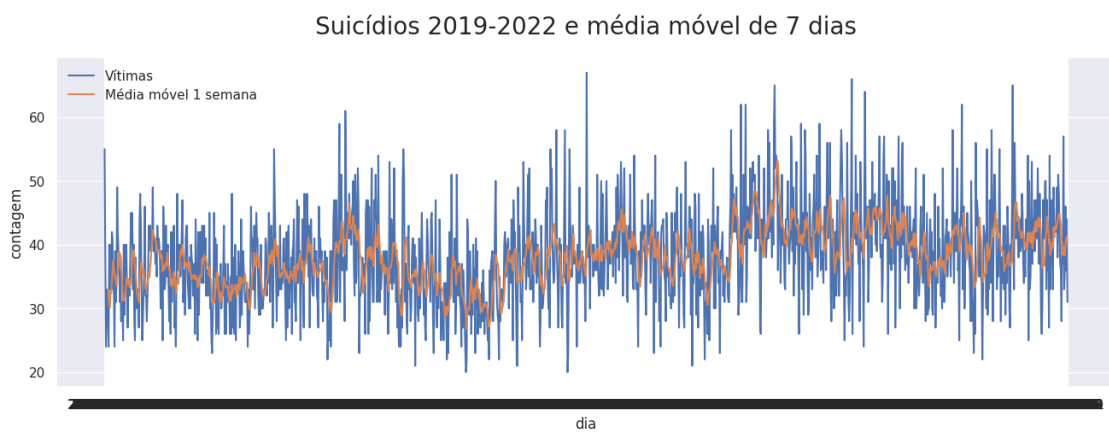
[ ]:
```
suicidio_por_data['media_movel_por_semana'] = suicidio_por_data.contagem.
↪rolling(7).mean()
suicidio_por_data
```

[ ]:

```
             dia  contagem  media_movel_por_semana
39    2019-01-01        55                     NaN
725   2019-01-02        38                     NaN
1432  2019-01-03        24                     NaN
1062  2019-01-04        33                     NaN
1280  2019-01-05        29                     NaN
…            …       …                       …
876   2022-12-27        36               38.285714
213   2022-12-28        46               39.714286
880   2022-12-29        36               40.857143
296   2022-12-30        44               41.285714
1210  2022-12-31        31               40.428571

[1461 rows x 3 columns]
```

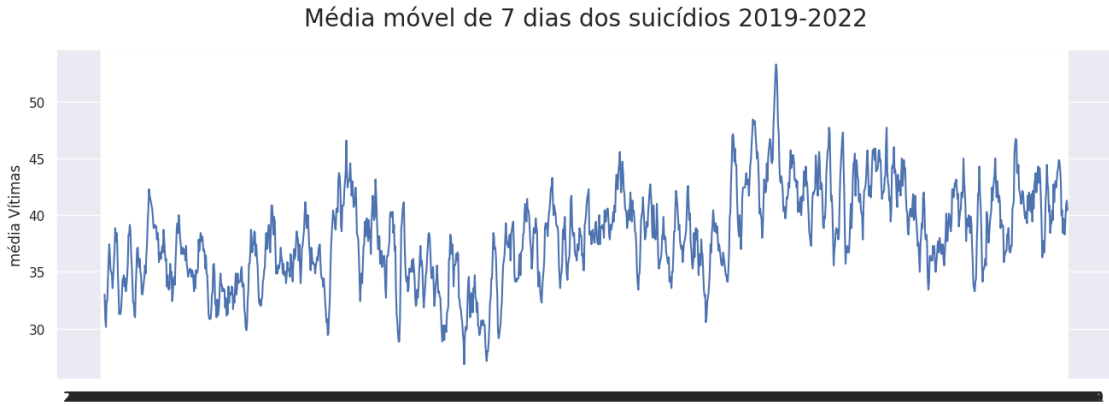
[ ]:

```
# gráfico média móvel
plt.figure(figsize = (16, 5))

sns.lineplot(x='dia', y='contagem', data=suicidio_por_data, label='Vítimas')
sns.lineplot(x='dia', y='media_movel_por_semana', data=suicidio_por_data,␣
 ↪label='Média móvel 1 semana')
plt.title("Suicídios 2019-2022 e média móvel de 7 dias", fontsize=20, pad=20)
plt.show()
```



[ ]:

```
plt.figure(figsize = (16, 5))

sns.lineplot(x='dia', y='media_movel_por_semana', data=suicidio_por_data)
plt.title("Média móvel de 7 dias dos suicídios 2019-2022", fontsize=20, pad=20)
plt.ylabel('média Vítimas')
plt.xlabel('')
plt.show()
```



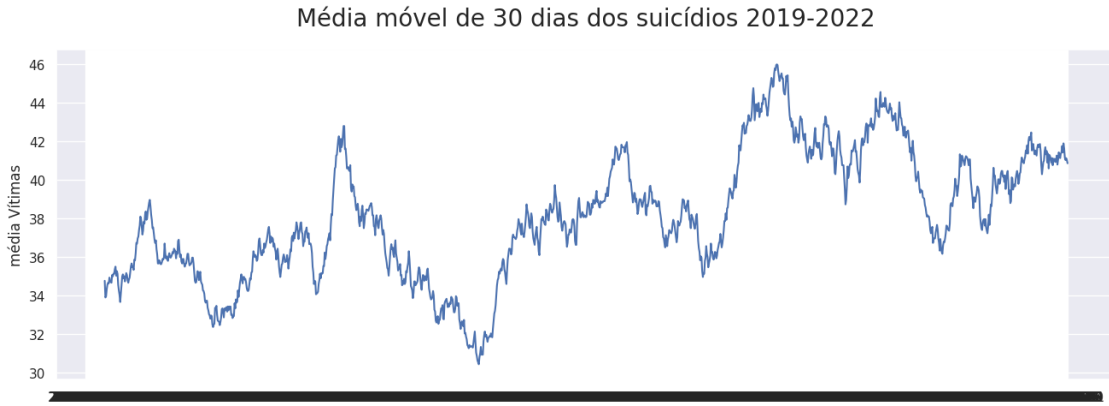
[ ]:
```
#média móvel para 30 dias
suicidio_por_data['media_movel_por_mes'] = suicidio_por_data.contagem.
 ↪rolling(30).mean()
suicidio_por_data
```

[ ]:

	dia	contagem	media_movel_por_semana	media_movel_por_mes
39	2019-01-01	55	NaN	NaN
725	2019-01-02	38	NaN	NaN
1432	2019-01-03	24	NaN	NaN
1062	2019-01-04	33	NaN	NaN
1280	2019-01-05	29	NaN	NaN
…	…	…	…	…
876	2022-12-27	36	38.285714	41.166667
213	2022-12-28	46	39.714286	41.033333
880	2022-12-29	36	40.857143	41.133333
296	2022-12-30	44	41.285714	41.033333
1210	2022-12-31	31	40.428571	40.866667

[1461 rows x 4 columns]

[ ]:
```
plt.figure(figsize = (16, 5))

sns.lineplot(x='dia', y='media_movel_por_mes', data=suicidio_por_data)
plt.title("Média móvel de 30 dias dos suicídios 2019-2022", fontsize=20, pad=20)
plt.ylabel('média Vítimas')
plt.xlabel('')
plt.show()
```



Considerando a média móvel de 30 dias, podemos observar um fato bastante interessante e curiosa: o número diminui nos meses de julho e tem tendência de crescimento para o início do ano. Particularmente, eu esperava que a distribuição fosse aleatória e, em média, constante durante o ano. Podemos criar 2 perguntas para estudar com mais atenção: * Qual é a relação entre o mês de julho e a queda de suicídios? * Qual a relação de crecimento e o final do ano?

OBSERVAÇÃO!

Considerando o Paradoxo de Simpson, faremos uma análise dos anos somados e depois individualmente por ano.

2.2 Suicídio por raça

[ ]:
```
# valores únicos e a prevalência
suicidio_por_raca = gerar_dataframe_categorico(dados.RACACOR, 'RACACOR',␣
↪'contagem')
suicidio_por_raca
```

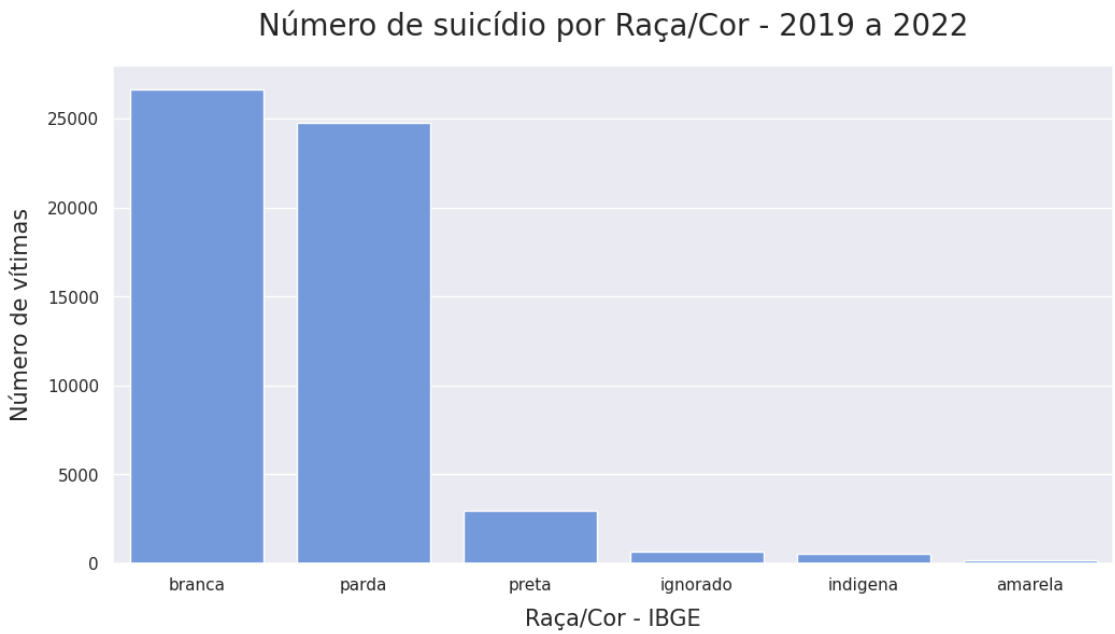
[ ]:
```
    RACACOR  contagem
0    branca     26642
1     parda     24743
2     preta      2966
3  ignorado       683
4  indigena       526
5   amarela       167
```

[ ]:
```
# valor percentual
dados.RACACOR.value_counts(normalize=True) * 100
```

[ ]:
```
branca      47.808064
parda       44.400380
preta        5.322375
```

```
ignorado     1.225618
indigena     0.943887
amarela      0.299675
Name: RACACOR, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('RACACOR', 'contagem', suicidio_por_raca,
                   'Número de suicídio por Raça/Cor - 2019 a 2022', 'Raça/Cor -␣
↪IBGE', 'Número de vítimas')
```



O que podemos observar é que quase metade dos suicídios foram de brancos (47,8%), seguidos de perto pelos pardos (44,4%). Podemos observar que a incidência para pretos é proporcionalmente baixa (5,3%).

[ ]:
```
# valores únicos e a prevalência
suicidio_por_raca_19 = gerar_dataframe_categorico(dados_2019.RACACOR,␣
↪'RACACOR', 'contagem')
suicidio_por_raca_19
```

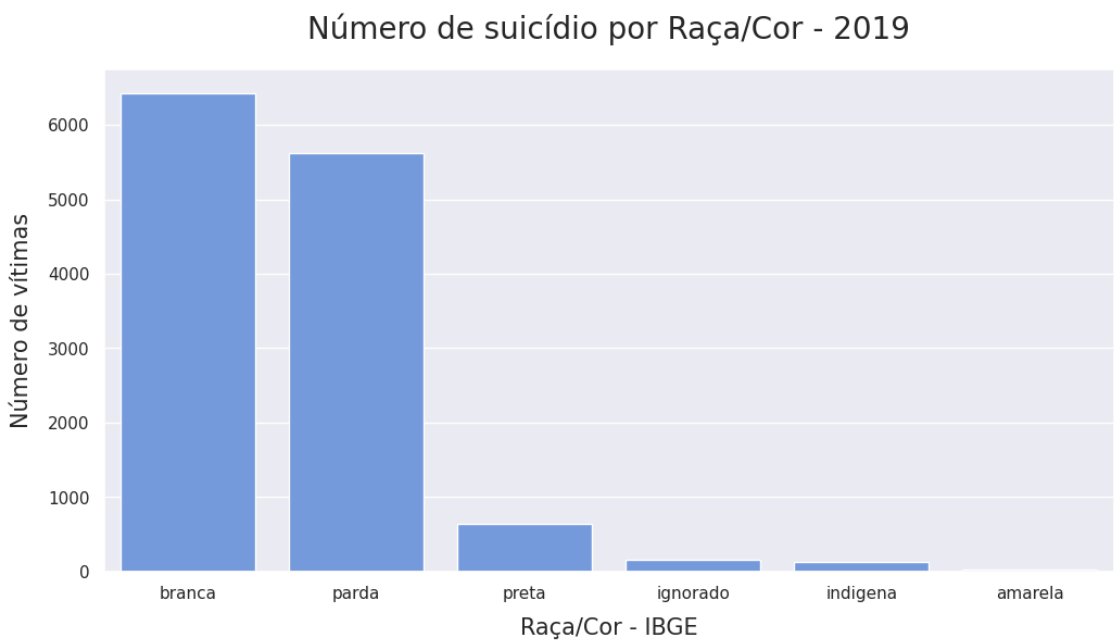
[ ]:

	RACACOR	contagem
0	branca	6425
1	parda	5616
2	preta	647
3	ignorado	166
4	indigena	127
5	amarela	28

[ ]:
```
# valor percentual
dados_2019.RACACOR.value_counts(normalize=True) * 100
```

[ ]:
```
branca      49.388885
parda       43.170113
preta        4.973480
ignorado     1.276040
indigena     0.976247
amarela      0.215236
Name: RACACOR, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('RACACOR', 'contagem', suicidio_por_raca_19,
                   'Número de suicídio por Raça/Cor - 2019', 'Raça/Cor - IBGE',␣
 ↪'Número de vítimas')
```



[ ]:
```
suicidio_por_raca_20 = gerar_dataframe_categorico(dados_2020.RACACOR,␣
 ↪'RACACOR', 'contagem')
suicidio_por_raca_20
```

[ ]:

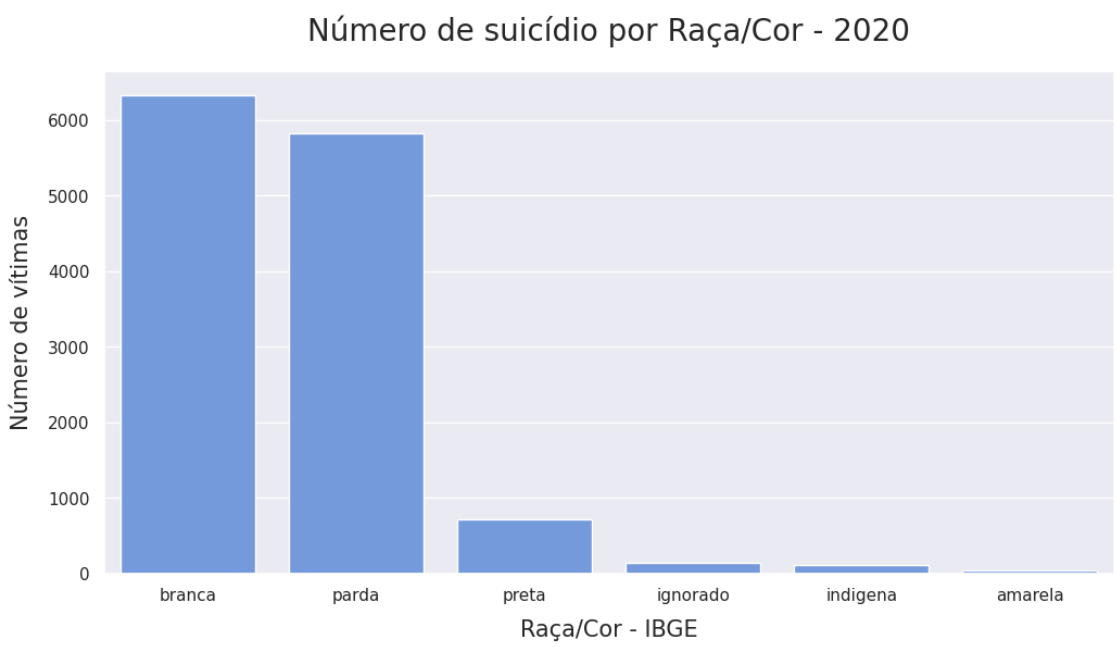
	RACACOR	contagem
0	branca	6327
1	parda	5816
2	preta	720
3	ignorado	141
4	indigena	115

```
5   amarela        47
```

[ ]:
```python
# valor percentual
dados_2020.RACACOR.value_counts(normalize=True) * 100
```

[ ]:
```
branca      48.055598
parda       44.174389
preta        5.468631
ignorado     1.070940
indigena     0.873462
amarela      0.356980
Name: RACACOR, dtype: float64
```

[ ]:
```python
#gráfico de colunas conforme função personalizada
grafico_de_colunas('RACACOR', 'contagem', suicidio_por_raca_20,
                   'Número de suicídio por Raça/Cor - 2020', 'Raça/Cor - IBGE',␣
 ↪'Número de vítimas')
```



[ ]:
```python
# valores únicos e a prevalência
suicidio_por_raca_21 = gerar_dataframe_categorico(dados_2021.RACACOR,␣
 ↪'RACACOR', 'contagem')
suicidio_por_raca_21
```

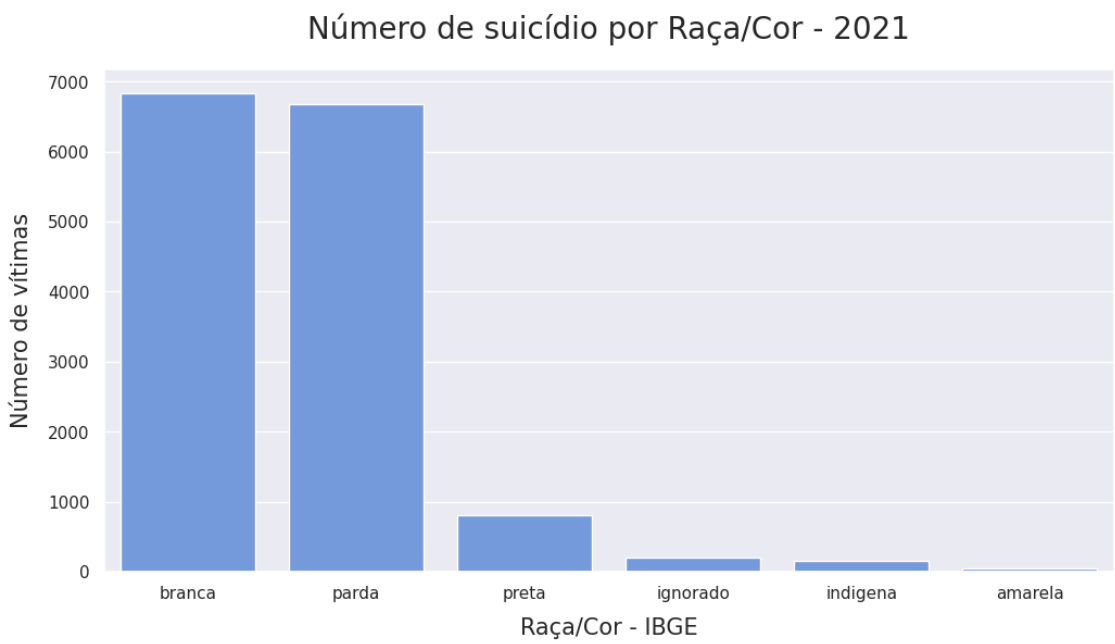
[ ]:
```
    RACACOR  contagem
0    branca      6834
1     parda      6681
```

```
2     preta       806
3  ignorado       208
4  indigena       151
5   amarela        42
```

[ ]:
```python
# valor percentual
dados_2021.RACACOR.value_counts(normalize=True) * 100
```

[ ]:
```
branca      46.420323
parda       45.381062
preta        5.474800
ignorado     1.412852
indigena     1.025676
amarela      0.285287
Name: RACACOR, dtype: float64
```

[ ]:
```python
#gráfico de colunas conforme função personalizada
grafico_de_colunas('RACACOR', 'contagem', suicidio_por_raca_21,
                   'Número de suicídio por Raça/Cor - 2021', 'Raça/Cor - IBGE',␣
 ↪'Número de vítimas')
```



[ ]:
```python
# valores únicos e a prevalência
suicidio_por_raca_22 = gerar_dataframe_categorico(dados_2022.RACACOR,␣
 ↪'RACACOR', 'contagem')
suicidio_por_raca_22
```

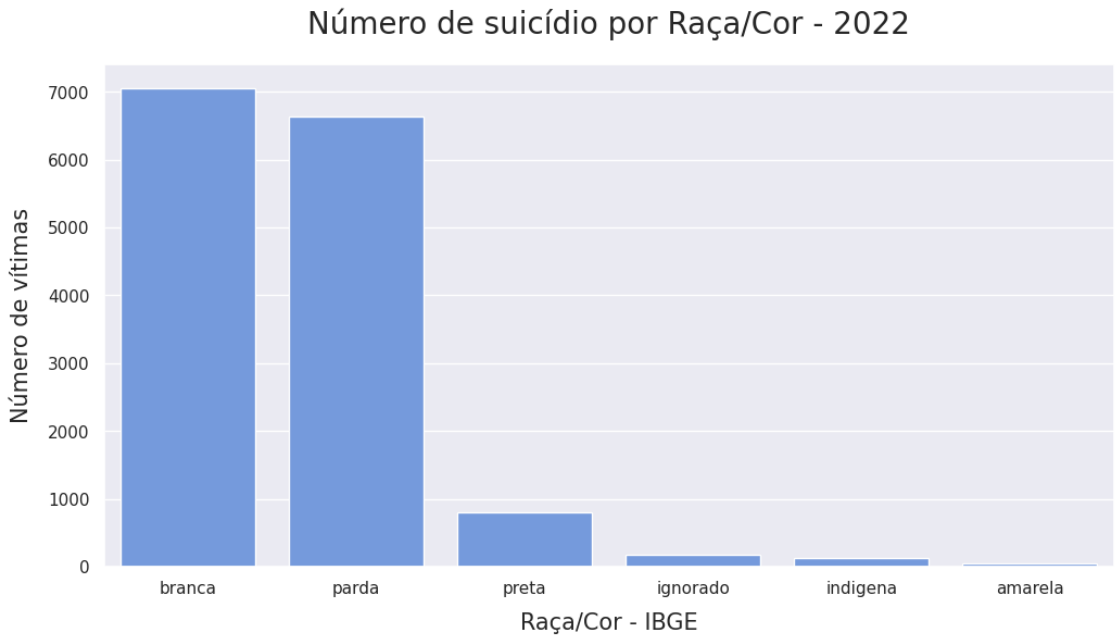
[ ]:

	RACACOR	contagem
0	branca	7056
1	parda	6630
2	preta	793
3	ignorado	168
4	indigena	133
5	amarela	50

[ ]:
```
# valor percentual
dados_2022.RACACOR.value_counts(normalize=True) * 100
```

[ ]:
```
branca      47.579231
parda       44.706676
preta        5.347269
ignorado     1.132839
indigena     0.896831
amarela      0.337154
Name: RACACOR, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('RACACOR', 'contagem', suicidio_por_raca_22,
                   'Número de suicídio por Raça/Cor - 2022', 'Raça/Cor - IBGE',␣
 ↪'Número de vítimas')
```



Podemos observar que as distribuições pelos anos é comparável ao somatório dos anos.

2.3 Suicídio por Sexo

[ ]:
```
suicidio_por_sexo = gerar_dataframe_categorico(dados.SEXO, 'SEXO', 'contagem')
suicidio_por_sexo
```

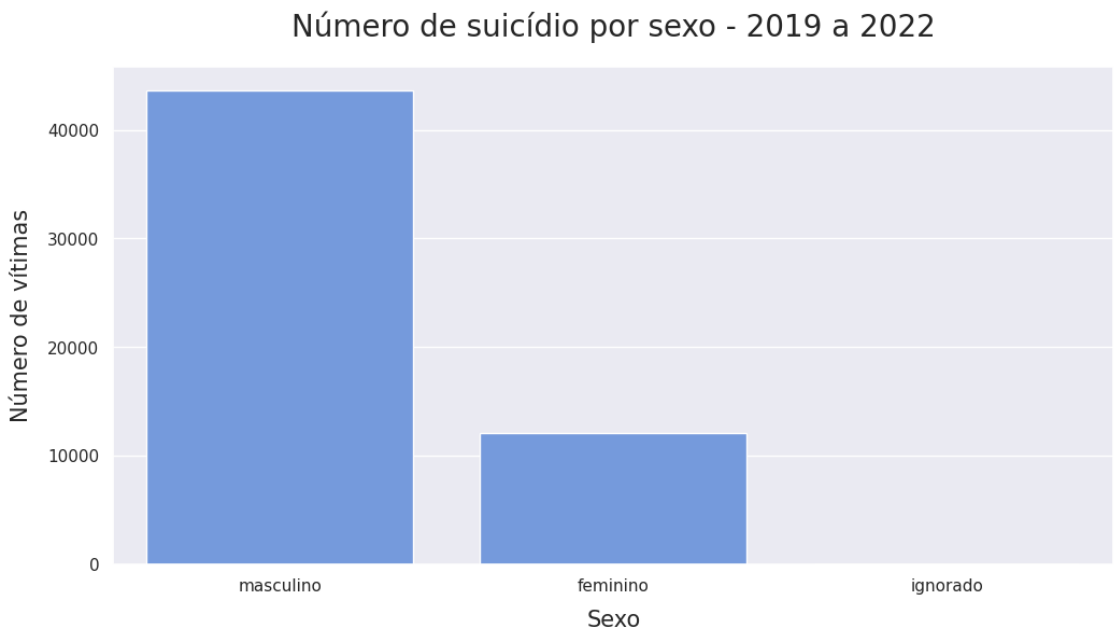
[ ]:

	SEXO	contagem
0	masculino	43633
1	feminino	12084
2	ignorado	10

[ ]:
```
# percentual
dados.SEXO.value_counts(normalize=True) *100
```

[ ]:
```
masculino    78.297773
feminino     21.684282
ignorado      0.017945
Name: SEXO, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('SEXO', 'contagem', suicidio_por_sexo,
                   'Número de suicídio por sexo - 2019 a 2022', 'Sexo', 'Número␣
 ↪de vítimas')
```



Podemos observar que a diferença entre homens e mulheres é bastante relevante. Enquanto os homens têm mais de 40 mil casos (78%), é pouco mais de 10 mil o caso de mulheres (21,68%). Infelizmente não existe a possibilidade de fazer análise quanto às identidades de gêneros.

[ ]:
```
suicidio_por_sexo_19 = gerar_dataframe_categorico(dados_2019.SEXO, 'SEXO',␣
↪'contagem')
suicidio_por_sexo_19
```

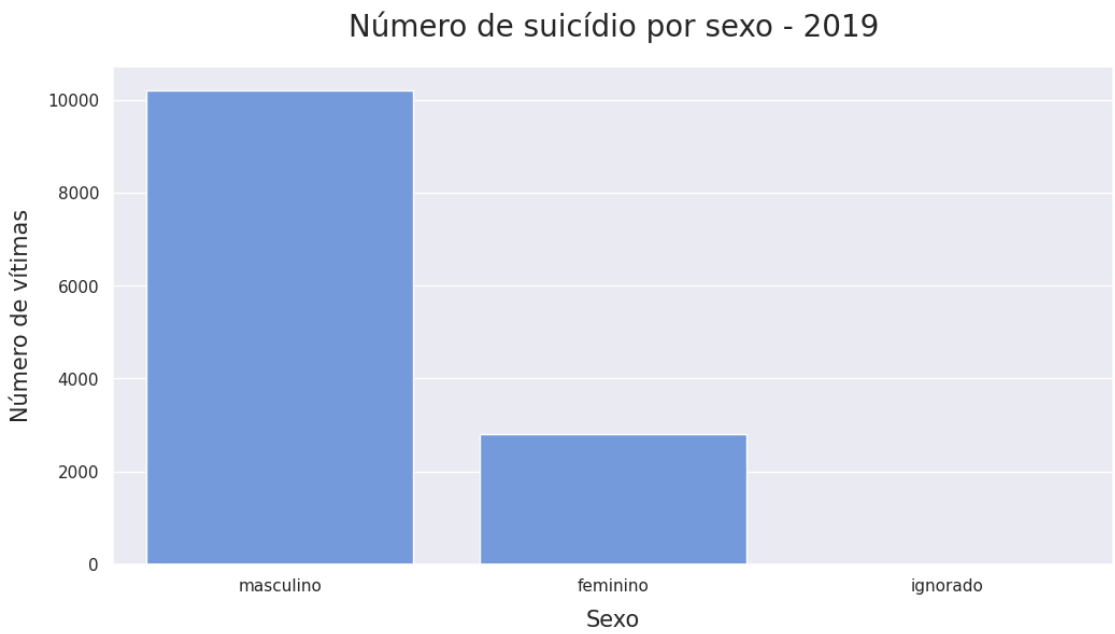
[ ]:

	SEXO	contagem
0	masculino	10202
1	feminino	2806
2	ignorado	1

[ ]:
```
# percentual
dados_2019.SEXO.value_counts(normalize=True) *100
```

[ ]:
```
masculino    78.422630
feminino     21.569683
ignorado      0.007687
Name: SEXO, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('SEXO', 'contagem', suicidio_por_sexo_19,
                   'Número de suicídio por sexo - 2019', 'Sexo', 'Número de␣
↪vítimas')
```



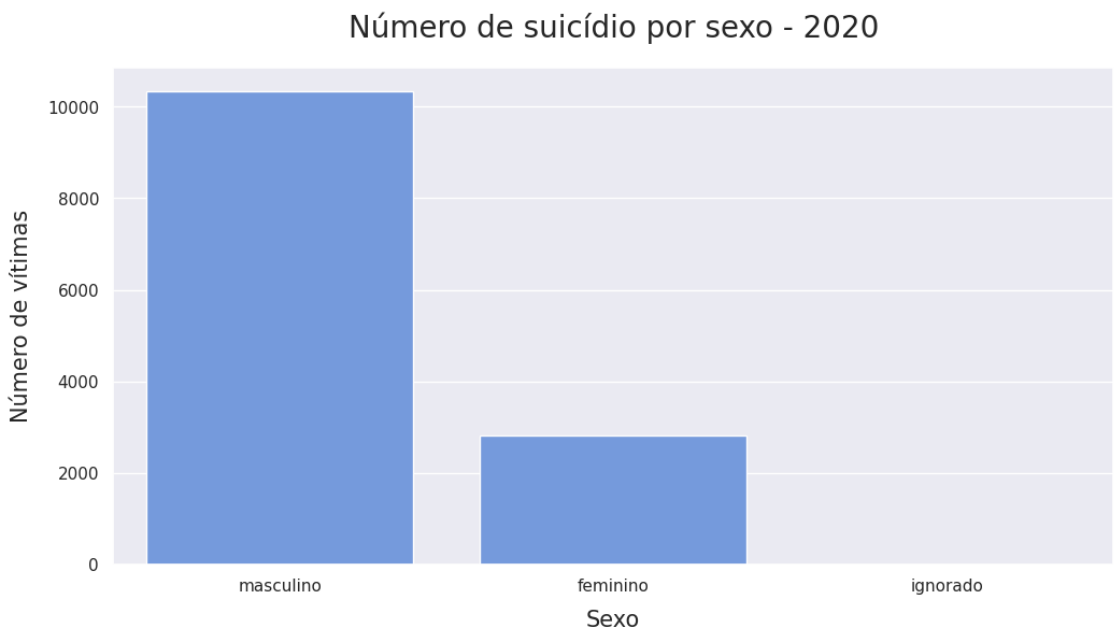
[ ]:
```
suicidio_por_sexo_20 = gerar_dataframe_categorico(dados_2020.SEXO, 'SEXO',␣
↪'contagem')
suicidio_por_sexo_20
```

[ ]:
```
        SEXO  contagem
0  masculino     10340
1   feminino      2823
2   ignorado         3
```

[ ]:
```python
# percentual
dados_2020.SEXO.value_counts(normalize=True) *100
```

[ ]:
```
masculino    78.535622
feminino     21.441592
ignorado      0.022786
Name: SEXO, dtype: float64
```

[ ]:
```python
#gráfico de colunas conforme função personalizada
grafico_de_colunas('SEXO', 'contagem', suicidio_por_sexo_20,
                   'Número de suicídio por sexo - 2020', 'Sexo', 'Número de␣
 ↪vítimas')
```



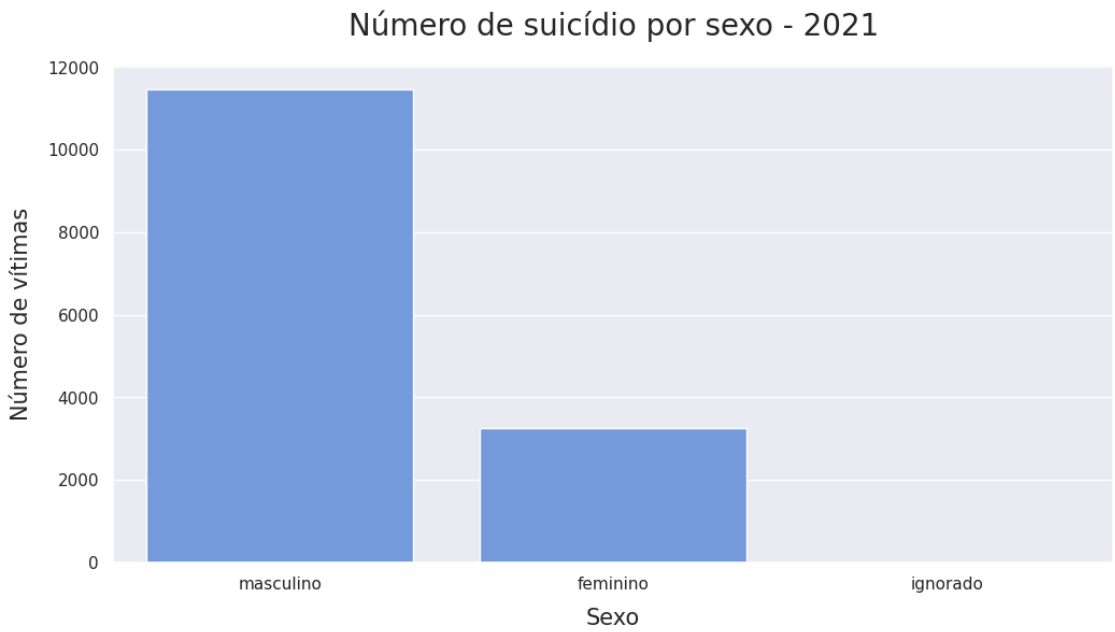
[ ]:
```python
suicidio_por_sexo_21 = gerar_dataframe_categorico(dados_2021.SEXO, 'SEXO',␣
 ↪'contagem')
suicidio_por_sexo_21
```

[ ]:
```
        SEXO  contagem
0  masculino     11460
1   feminino      3258
2   ignorado         4
```

[ ]:
```
# percentual
dados_2021.SEXO.value_counts(normalize=True) *100
```

[ ]:
```
masculino    77.842684
feminino     22.130145
ignorado      0.027170
Name: SEXO, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('SEXO', 'contagem', suicidio_por_sexo_21,
                   'Número de suicídio por sexo - 2021', 'Sexo', 'Número de␣
 ↪vítimas')
```



[ ]:
```
suicidio_por_sexo_22 = gerar_dataframe_categorico(dados_2022.SEXO, 'SEXO',␣
 ↪'contagem')
suicidio_por_sexo_22
```

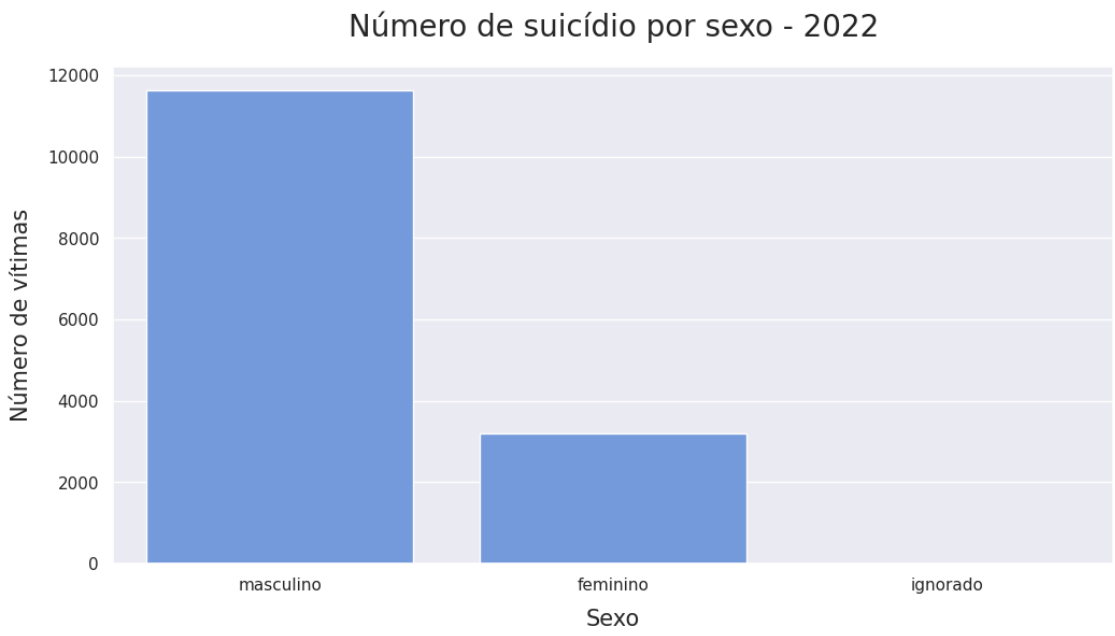
[ ]:

	SEXO	contagem
0	masculino	11631
1	feminino	3197
2	ignorado	2

[ ]:
```
# percentual
dados_2022.SEXO.value_counts(normalize=True) *100
```

[ ]:
```
masculino    78.428860
feminino     21.557653
ignorado      0.013486
Name: SEXO, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('SEXO', 'contagem', suicidio_por_sexo_22,
                   'Número de suicídio por sexo - 2022', 'Sexo', 'Número de␣
 ↪vítimas')
```



Podemos comprovar que a distribuição de suicídios por ano são comparáveis entre si e com a distribuição com o somatório dos anos de 2019 a 2022

2.3.1 Cruzamento raça e sexo

[ ]:
```
dados[['RACACOR', 'SEXO', 'DTOBITO']].groupby(['SEXO', 'RACACOR']).count().
 ↪sort_values(by=['SEXO', 'DTOBITO']).unstack(fill_value=0)
```

[ ]:

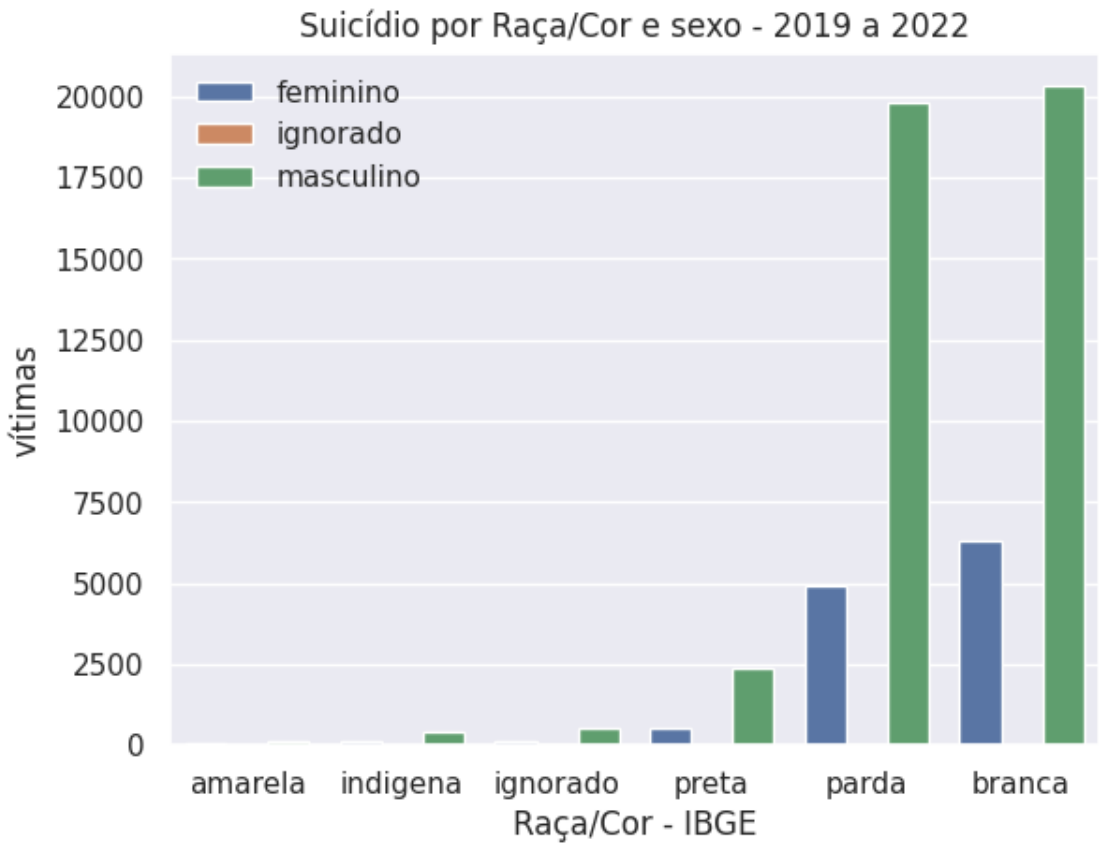
	DTOBITO					
RACACOR	amarela	branca	ignorado	indigena	parda	preta
SEXO						
feminino	45	6323	128	115	4911	562
ignorado	0	0	7	0	2	1
masculino	122	20319	548	411	19830	2403

[ ]:
```
raca_e_sexo = dados[['RACACOR', 'SEXO', 'DTOBITO']].groupby(['SEXO',␣
↪'RACACOR']).count().sort_values(by=['SEXO', 'DTOBITO']).reset_index()
raca_e_sexo
```

[ ]:

	SEXO	RACACOR	DTOBITO
0	feminino	amarela	45
1	feminino	indigena	115
2	feminino	ignorado	128
3	feminino	preta	562
4	feminino	parda	4911
5	feminino	branca	6323
6	ignorado	preta	1
7	ignorado	parda	2
8	ignorado	ignorado	7
9	masculino	amarela	122
10	masculino	indigena	411
11	masculino	ignorado	548
12	masculino	preta	2403
13	masculino	parda	19830
14	masculino	branca	20319

[ ]:
```
sns.barplot(x='RACACOR', y='DTOBITO', hue='SEXO', data=raca_e_sexo)
plt.title('Suicídio por Raça/Cor e sexo - 2019 a 2022')
plt.xlabel('Raça/Cor - IBGE')
plt.ylabel('vítimas')
plt.legend(loc='upper left')
plt.show()
```



Podemos comprovar que a distribuição de Sexo e Raça/cor se comportam da mesma maneira quando avaliadas separadamente.

2.4 Suicídio por escolaridade

[ ]:
```
dados.ESC.value_counts()
```

[ ]:
```
de 8 a 11 anos    17897
de 4 a 7 anos     12567
ignorado          11557
12 anos e mais     5846
de 1 a 3 anos      5635
nenhuma            2225
Name: ESC, dtype: int64
```

[ ]:
```
dados.ESC2010.value_counts()
```

[ ]:
```
médio (antigo 2º grau)            13191
fundamental II (5ª a 8ª série)    12951
ignorado                          11751
```

```
fundamental I (1ª a 4ª série)      9817
superior completo                  3872
sem escolaridade                   2205
superior incompleto                1940
Name: ESC2010, dtype: int64
```

[ ]:
```
dados.ESCFALAGR1.value_counts()
```

[ ]:
```
ignorado                      11751
fundamental II incompleto      8285
ensino médio incompleto        6725
fundamental I incompleto       5593
sem escolaridade               5427
ensino médio completo          5184
fundamental II completo        4666
fundamental I completo         4224
superior completo              3872
Name: ESCFALAGR1, dtype: int64
```

A metodologia de contabilização de escolaridade é confusa e com os números inconsistentes entre si, contudo as variaveis ESC2010 e ESCFALAGR1 têm valores comuns e dentre eles escolherei o ESCFALAGR1, pois apresenta maior detalhamento.

[ ]:
```
suicidio_por_escolaridade = gerar_dataframe_categorico(dados.ESCFALAGR1,␣
↪'escolaridade', 'contagem')
suicidio_por_escolaridade
```

[ ]:

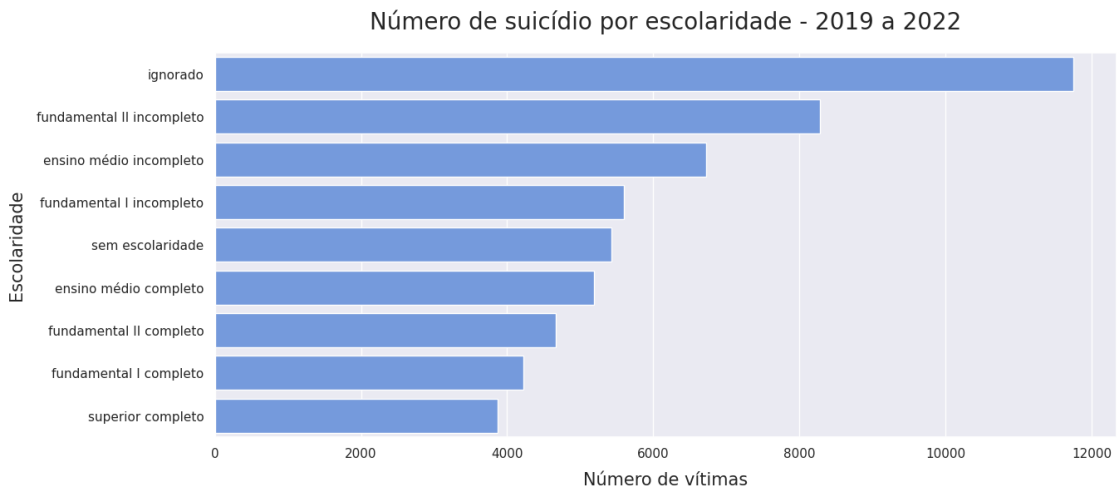
	escolaridade	contagem
0	ignorado	11751
1	fundamental II incompleto	8285
2	ensino médio incompleto	6725
3	fundamental I incompleto	5593
4	sem escolaridade	5427
5	ensino médio completo	5184
6	fundamental II completo	4666
7	fundamental I completo	4224
8	superior completo	3872

[ ]:
```
dados.ESCFALAGR1.value_counts(normalize=True)*100
```

[ ]:
```
ignorado                      21.086726
fundamental II incompleto     14.867120
ensino médio incompleto       12.067759
fundamental I incompleto      10.036428
sem escolaridade               9.738547
ensino médio completo          9.302493
fundamental II completo        8.372961
```

```
fundamental I completo        7.579809
superior completo             6.948158
Name: ESCFALAGR1, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'escolaridade', suicidio_por_escolaridade,
                    'Número de suicídio por escolaridade - 2019 a 2022', 'Número␣
 ↪de vítimas', 'Escolaridade')
```



Excluindo a parcela de escolaridade ignorada, Podemos observar que a distribuição de suicídios é maior na parcela da população com menor escolaridade.

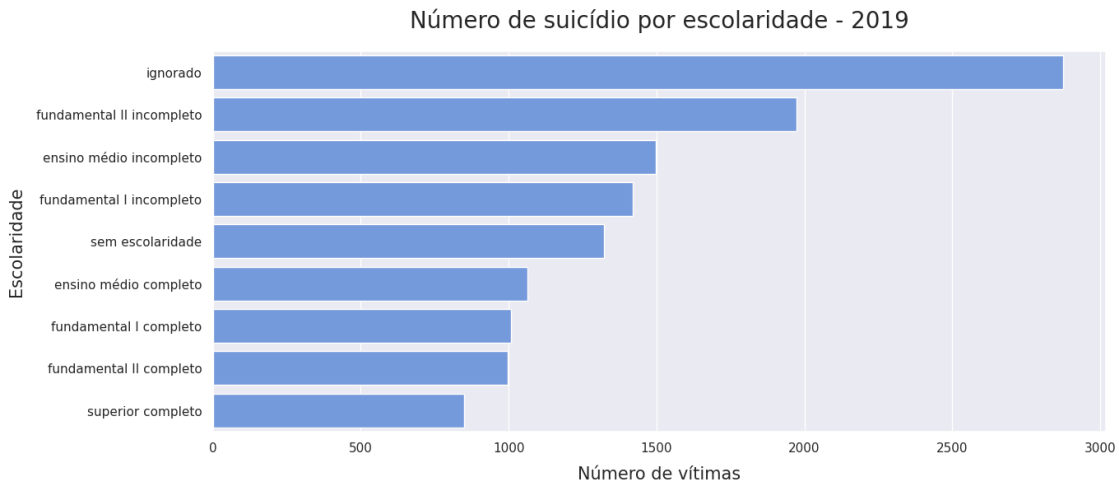
[ ]:
```
suicidio_por_escolaridade_19 = gerar_dataframe_categorico(dados_2019.
 ↪ESCFALAGR1, 'escolaridade', 'contagem')
suicidio_por_escolaridade_19
```

[ ]:

	escolaridade	contagem
0	ignorado	2875
1	fundamental II incompleto	1973
2	ensino médio incompleto	1498
3	fundamental I incompleto	1419
4	sem escolaridade	1323
5	ensino médio completo	1065
6	fundamental I completo	1009
7	fundamental II completo	998
8	superior completo	849

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'escolaridade', suicidio_por_escolaridade_19,
```

```
                    'Número de suicídio por escolaridade - 2019','Número de␣
↪vítimas', 'Escolaridade')
```

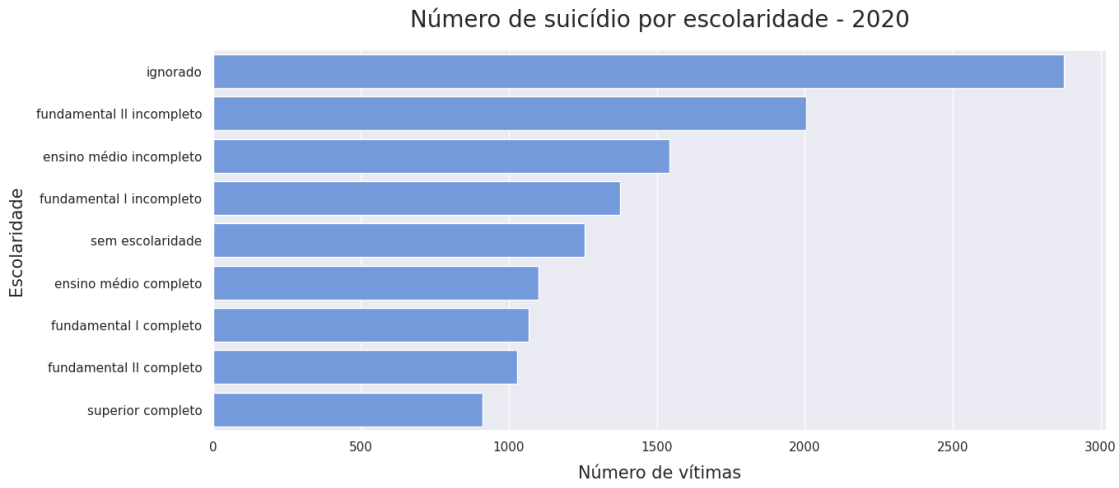


[ ]:
```
suicidio_por_escolaridade_20 = gerar_dataframe_categorico(dados_2020.
↪ESCFALAGR1, 'escolaridade', 'contagem')
suicidio_por_escolaridade_20
```

[ ]:

	escolaridade	contagem
0	ignorado	2876
1	fundamental II incompleto	2006
2	ensino médio incompleto	1543
3	fundamental I incompleto	1377
4	sem escolaridade	1257
5	ensino médio completo	1101
6	fundamental I completo	1066
7	fundamental II completo	1029
8	superior completo	911

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'escolaridade', suicidio_por_escolaridade_20,
                    'Número de suicídio por escolaridade - 2020','Número de␣
↪vítimas', 'Escolaridade')
```

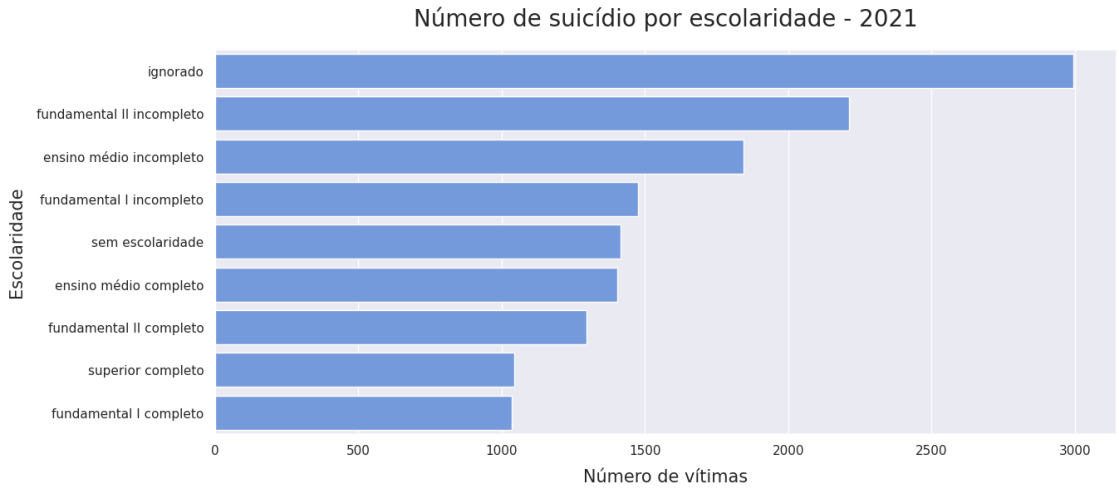


[ ]:
```
suicidio_por_escolaridade_21 = gerar_dataframe_categorico(dados_2021.
↪ESCFALAGR1, 'escolaridade', 'contagem')
suicidio_por_escolaridade_21
```

[ ]:

	escolaridade	contagem
0	ignorado	2995
1	fundamental II incompleto	2212
2	ensino médio incompleto	1844
3	fundamental I incompleto	1475
4	sem escolaridade	1416
5	ensino médio completo	1405
6	fundamental II completo	1297
7	superior completo	1043
8	fundamental I completo	1035

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'escolaridade', suicidio_por_escolaridade_21,
                  'Número de suicídio por escolaridade - 2021', 'Número de␣
↪vítimas', 'Escolaridade')
```

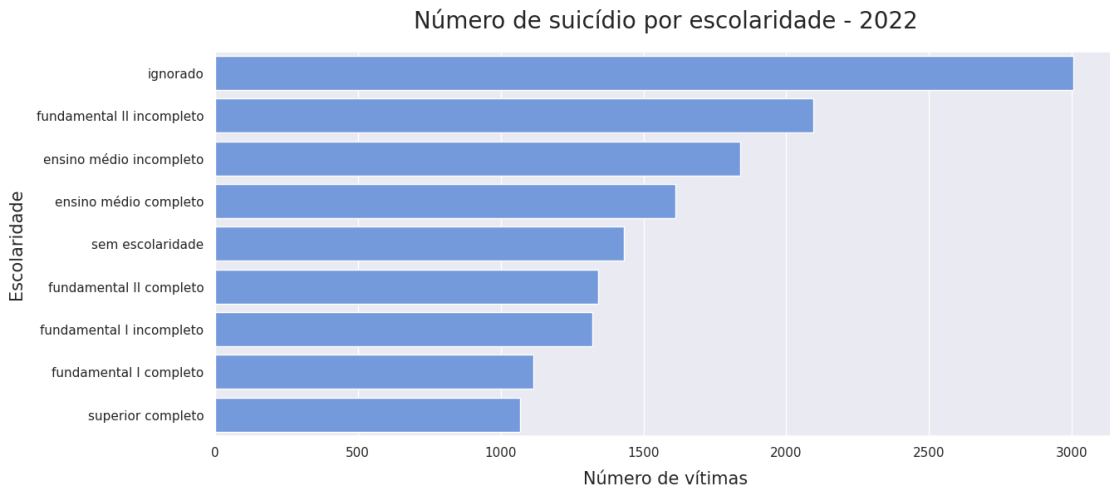


[ ]:
```
suicidio_por_escolaridade_22 = gerar_dataframe_categorico(dados_2022.
↪ESCFALAGR1, 'escolaridade', 'contagem')
suicidio_por_escolaridade_22
```

[ ]:

	escolaridade	contagem
0	ignorado	3005
1	fundamental II incompleto	2094
2	ensino médio incompleto	1840
3	ensino médio completo	1613
4	sem escolaridade	1431
5	fundamental II completo	1342
6	fundamental I incompleto	1322
7	fundamental I completo	1114
8	superior completo	1069

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'escolaridade', suicidio_por_escolaridade_22,
                  'Número de suicídio por escolaridade - 2022', 'Número de␣
↪vítimas', 'Escolaridade')
```



Podemos confirmar que as distribuições de escolaridade são comparáveis entre os anos e entre a consolidação dos anos.

2.4.1 Escolaridade por sexo

[ ]:
```
dados[['ESCFALAGR1', 'SEXO', 'DTOBITO']].groupby(['SEXO', 'ESCFALAGR1']).
 ↪count().sort_values(by=['SEXO', 'DTOBITO']).unstack(fill_value=0)
```

[ ]:
```
                          DTOBITO                                  \
ESCFALAGR1 ensino médio completo ensino médio incompleto
SEXO
feminino                    1294                    1724
ignorado                       0                       1
masculino                   3890                    5000

                                                                   \
ESCFALAGR1 fundamental I completo fundamental I incompleto
SEXO
feminino                     732                     902
ignorado                       0                       0
masculino                   3492                    4691

                                                                                  \
ESCFALAGR1 fundamental II completo fundamental II incompleto ignorado
SEXO
feminino                       939                      1607     2300
ignorado                         0                         0        9
masculino                     3727                      6678     9442
```

```
ESCFALAGR1 sem escolaridade superior completo
SEXO
feminino                1285              1301
ignorado                   0                 0
masculino               4142              2571
```

Há de se destacar que a diferença entre homens e mulheres é muito grande (diferença mínima de 200%) em quase todas as escolaridades, exceto com o nível superior completo (97%)

[ ]:
```
escolaridade_por_sexo = dados[['ESCFALAGR1', 'SEXO', 'DTOBITO']].
↪groupby(['SEXO', 'ESCFALAGR1']).count().sort_values(by=['SEXO', 'DTOBITO']).
↪reset_index()
escolaridade_por_sexo
```

[ ]:

	SEXO	ESCFALAGR1	DTOBITO
0	feminino	fundamental I completo	732
1	feminino	fundamental I incompleto	902
2	feminino	fundamental II completo	939
3	feminino	sem escolaridade	1285
4	feminino	ensino médio completo	1294
5	feminino	superior completo	1301
6	feminino	fundamental II incompleto	1607
7	feminino	ensino médio incompleto	1724
8	feminino	ignorado	2300
9	ignorado	ensino médio incompleto	1
10	ignorado	ignorado	9
11	masculino	superior completo	2571
12	masculino	fundamental I completo	3492
13	masculino	fundamental II completo	3727
14	masculino	ensino médio completo	3890
15	masculino	sem escolaridade	4142
16	masculino	fundamental I incompleto	4691
17	masculino	ensino médio incompleto	5000
18	masculino	fundamental II incompleto	6678
19	masculino	ignorado	9442

[ ]:
```
sns.barplot(y='ESCFALAGR1', x='DTOBITO', hue='SEXO', data=escolaridade_por_sexo)
plt.title('Suicídio por escolaridade e sexo - 2019 a 2022')
plt.ylabel('Escolaridade')
plt.xlabel('vítimas')
plt.legend(loc='upper right')
plt.show()
```

2.5 Suicídio por hora

[ ]:
```
dados.HORAOBITO.value_counts()
```

[ ]:
```
08:01:00    2975
10:01:00    2869
17:01:00    2860
09:01:00    2824
18:01:00    2798
12:01:00    2769
16:01:00    2748
07:01:00    2722
11:01:00    2714
15:01:00    2653
14:01:00    2598
06:01:00    2594
19:01:00    2542
13:01:00    2451
20:01:00    2357
21:01:00    2117
22:01:00    2035
00:01:00    2018
23:01:00    1897
05:01:00    1855
```

```
01:01:00    1468
03:01:00    1341
04:01:00    1269
02:01:00    1253
Name: HORAOBITO, dtype: int64
```

Com exceção das 18 horas, há uma preferência alta para o dia em detrimento da noite ou madrugada.

[ ]:
```
dados_2019.HORAOBITO.value_counts()
```

[ ]:
```
08:01:00    685
09:01:00    671
16:01:00    666
12:01:00    664
18:01:00    653
07:01:00    646
10:01:00    644
17:01:00    640
06:01:00    630
15:01:00    618
11:01:00    612
19:01:00    598
14:01:00    597
20:01:00    554
13:01:00    550
21:01:00    502
23:01:00    481
00:01:00    461
22:01:00    458
05:01:00    435
01:01:00    331
03:01:00    320
02:01:00    309
04:01:00    284
Name: HORAOBITO, dtype: int64
```

[ ]:
```
dados_2020.HORAOBITO.value_counts()
```

[ ]:
```
09:01:00    730
08:01:00    691
17:01:00    677
11:01:00    665
12:01:00    659
10:01:00    643
07:01:00    639
16:01:00    634
19:01:00    630
```

```
18:01:00    625
15:01:00    611
06:01:00    604
14:01:00    583
13:01:00    568
20:01:00    552
21:01:00    491
22:01:00    482
00:01:00    477
05:01:00    462
23:01:00    459
01:01:00    348
03:01:00    316
02:01:00    313
04:01:00    307
Name: HORAOBITO, dtype: int64
```

[ ]:
```
dados_2021.HORAOBITO.value_counts()
```

[ ]:
```
08:01:00    801
10:01:00    787
17:01:00    762
18:01:00    756
16:01:00    745
12:01:00    727
11:01:00    717
07:01:00    717
06:01:00    714
14:01:00    697
15:01:00    696
09:01:00    678
13:01:00    668
19:01:00    644
20:01:00    626
21:01:00    576
00:01:00    557
22:01:00    527
23:01:00    485
05:01:00    477
01:01:00    385
03:01:00    342
04:01:00    330
02:01:00    308
Name: HORAOBITO, dtype: int64
```

[ ]:
```
dados_2022.HORAOBITO.value_counts()
```

[ ]:
```
08:01:00    798
10:01:00    795
17:01:00    781
18:01:00    764
09:01:00    745
15:01:00    728
14:01:00    721
11:01:00    720
07:01:00    720
12:01:00    719
16:01:00    703
19:01:00    670
13:01:00    665
06:01:00    646
20:01:00    625
22:01:00    568
21:01:00    548
00:01:00    523
05:01:00    481
23:01:00    472
01:01:00    404
03:01:00    363
04:01:00    348
02:01:00    323
Name: HORAOBITO, dtype: int64
```

O mesmo fenômeno da distribuição de casos ser durante o dia ser maior que à noite permanece durante os anos.

2.6 Suicídio por UF

[ ]:
```
suicidio_por_uf = gerar_dataframe_categorico(dados.UFOCOR, 'UF', 'contagem')
suicidio_por_uf
```

[ ]:

	UF	contagem
0	São Paulo	9614
1	Minas Gerais	6739
2	Rio Grande do Sul	5855
3	Paraná	3815
4	Santa Catarina	3291
5	Bahia	2959
6	Ceará	2725
7	Rio de Janeiro	2566
8	Goiás	2449
9	Pernambuco	1824
10	Pará	1550
11	Piauí	1342

```
12             Maranhão      1265
13             Amazonas      1144
14   Mato Grosso do Sul      1106
15              Paraíba      1052
16          Mato Grosso       985
17       Espírito Santo       939
18  Rio Grande do Norte       871
19     Distrito Federal       788
20              Alagoas       591
21             Rondônia       556
22            Tocantins       541
23              Sergipe       526
24                 Acre       273
25                Amapá       213
26              Roraima       148
```

[ ]:
```python
pd.DataFrame(dados.UFOCOR.value_counts(normalize=True)) *100
```

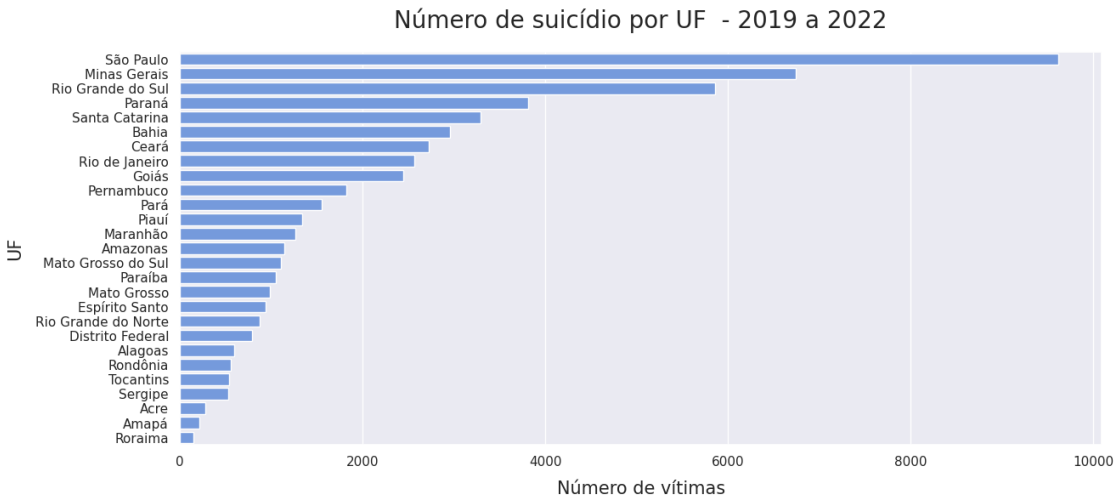
[ ]:
```
                        UFOCOR
São Paulo            17.251960
Minas Gerais         12.092881
Rio Grande do Sul    10.506577
Paraná                6.845874
Santa Catarina        5.905575
Bahia                 5.309814
Ceará                 4.889910
Rio de Janeiro        4.604590
Goiás                 4.394638
Pernambuco            3.273099
Pará                  2.781417
Piauí                 2.408168
Maranhão              2.269995
Amazonas              2.052865
Mato Grosso do Sul    1.984675
Paraíba               1.887774
Mato Grosso           1.767545
Espírito Santo        1.685000
Rio Grande do Norte   1.562977
Distrito Federal      1.414036
Alagoas               1.060527
Rondônia              0.997721
Tocantins             0.970804
Sergipe               0.943887
Acre                  0.489888
Amapá                 0.382220
Roraima               0.265580
```

[ ]:
```
dados.UFOCOR.value_counts(normalize=True)[:5].sum() * 100
```

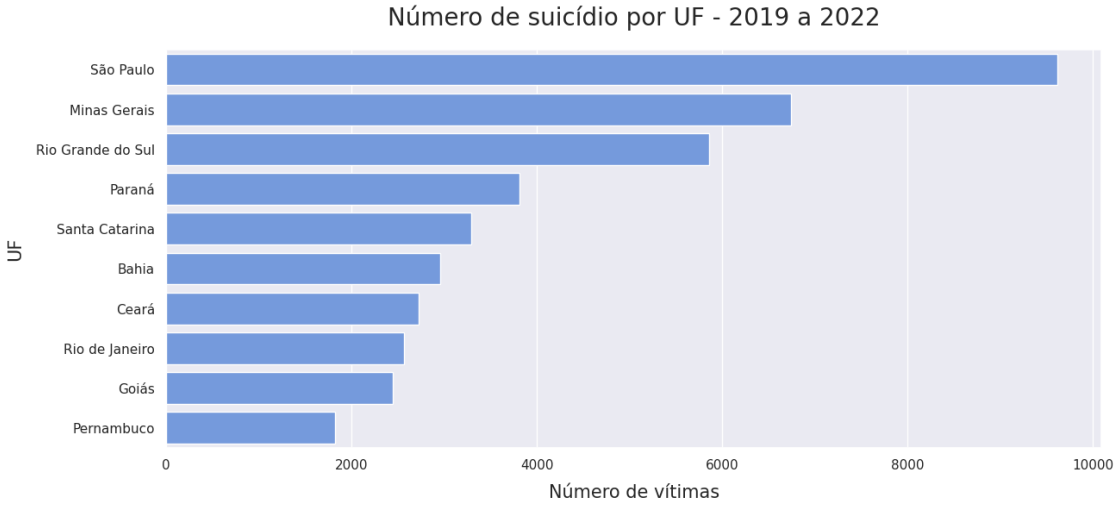
[ ]: 52.60286755073842

Os 5 primeiros estados correspondem por mais de 50% dos casos do país

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'UF', suicidio_por_uf,
                  'Número de suicídio por UF  - 2019 a 2022', 'Número de␣
 ↪vítimas', 'UF')
```



[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'UF', suicidio_por_uf.head(10),
                  'Número de suicídio por UF - 2019 a 2022','Número de␣
 ↪vítimas', 'UF')
```



[ ]:
```
suicidio_por_uf_19 = gerar_dataframe_categorico(dados_2019.UFOCOR, 'UF',␣
↪'contagem')
suicidio_por_uf_19
```

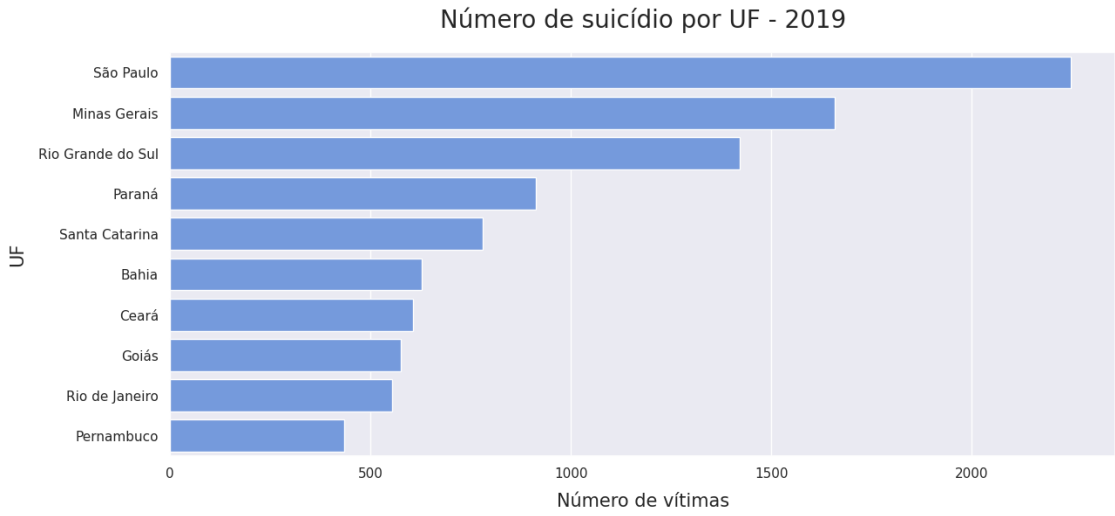
[ ]:

	UF	contagem
0	São Paulo	2246
1	Minas Gerais	1658
2	Rio Grande do Sul	1421
3	Paraná	913
4	Santa Catarina	780
5	Bahia	627
6	Ceará	606
7	Goiás	576
8	Rio de Janeiro	554
9	Pernambuco	434
10	Pará	330
11	Piauí	314
12	Maranhão	294
13	Mato Grosso do Sul	264
14	Amazonas	246
15	Espírito Santo	244
16	Paraíba	231
17	Mato Grosso	229
18	Distrito Federal	196
19	Rio Grande do Norte	191
20	Rondônia	134
21	Tocantins	126
22	Alagoas	120
23	Sergipe	111
24	Acre	67
25	Amapá	59
26	Roraima	38

[ ]:
```
dados_2019.UFOCOR.value_counts(normalize=True)[:5].sum() * 100
```

[ ]: 53.94726727650089

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'UF', suicidio_por_uf_19.head(10),
                  'Número de suicídio por UF - 2019', 'Número de vítimas',␣
↪'UF')
```




```
[ ]: suicidio_por_uf_20 = gerar_dataframe_categorico(dados_2020.UFOCOR, 'UF',␣
     ↪'contagem')
     suicidio_por_uf_20
```

[ ]:

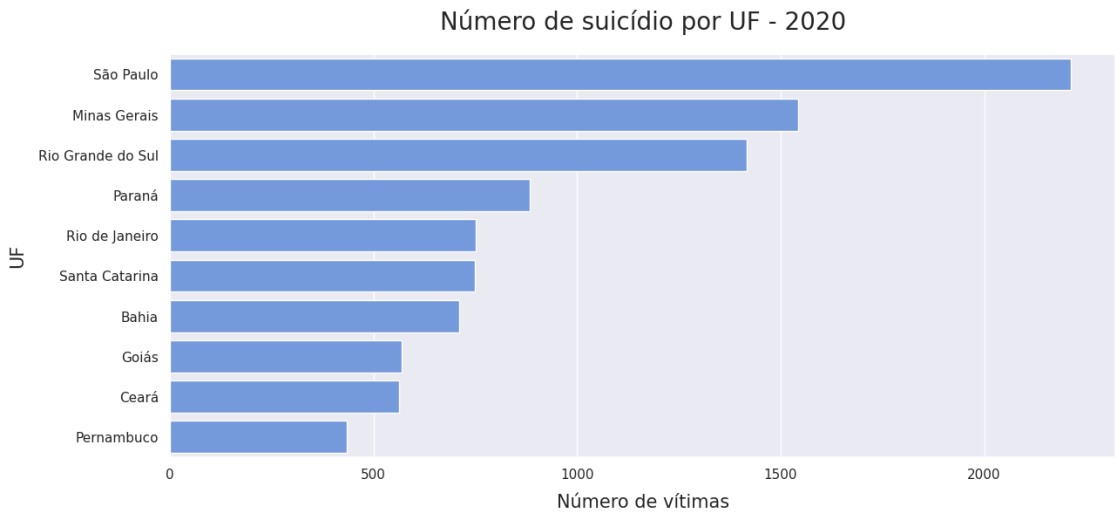
	UF	contagem
0	São Paulo	2212
1	Minas Gerais	1544
2	Rio Grande do Sul	1417
3	Paraná	883
4	Rio de Janeiro	750
5	Santa Catarina	749
6	Bahia	710
7	Goiás	568
8	Ceará	563
9	Pernambuco	434
10	Pará	372
11	Amazonas	313
12	Piauí	313
13	Maranhão	297
14	Paraíba	245
15	Espírito Santo	241
16	Mato Grosso do Sul	238
17	Mato Grosso	224
18	Rio Grande do Norte	218
19	Distrito Federal	196
20	Alagoas	162
21	Rondônia	135
22	Sergipe	129
23	Tocantins	111

```
24                 Acre        70
25                Amapá        45
26              Roraima        27
```

[ ]:
```
dados_2020.UFOCOR.value_counts(normalize=True)[:5].sum() * 100
```

[ ]: 51.69375664590612

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'UF', suicidio_por_uf_20.head(10),
                  'Número de suicídio por UF - 2020', 'Número de vítimas',␣
 ↪'UF')
```



[ ]:
```
suicidio_por_uf_21 = gerar_dataframe_categorico(dados_2021.UFOCOR, 'UF',␣
 ↪'contagem')
suicidio_por_uf_21
```

[ ]:

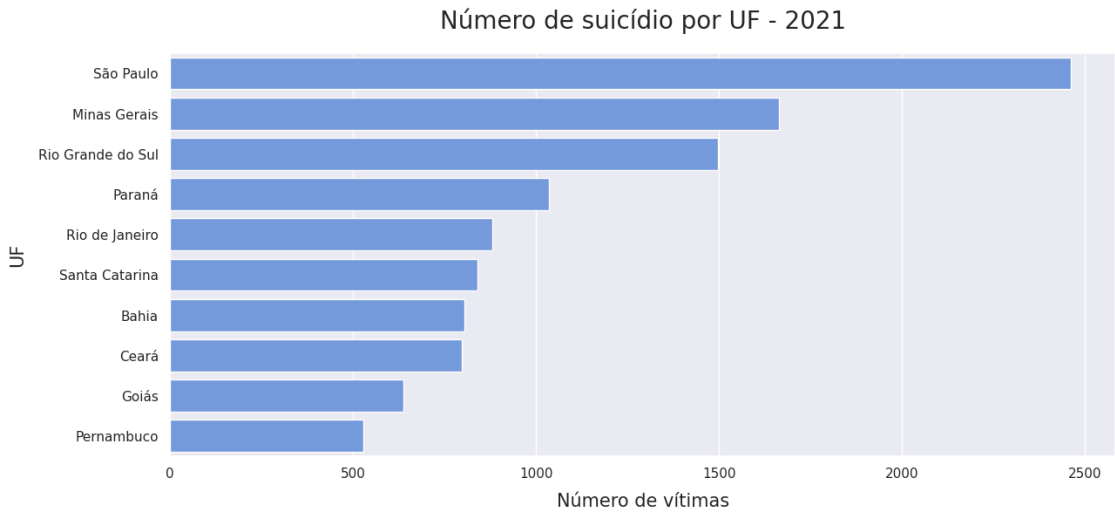
	UF	contagem
0	São Paulo	2460
1	Minas Gerais	1664
2	Rio Grande do Sul	1497
3	Paraná	1035
4	Rio de Janeiro	880
5	Santa Catarina	841
6	Bahia	804
7	Ceará	796
8	Goiás	637
9	Pernambuco	529
10	Pará	384

```
11                Piauí       365
12            Maranhão       333
13   Mato Grosso do Sul       297
14             Amazonas       294
15              Paraíba       264
16          Mato Grosso       246
17  Rio Grande do Norte       241
18       Espírito Santo       230
19     Distrito Federal       188
20            Tocantins       158
21              Alagoas       149
22             Rondônia       144
23              Sergipe       128
24                 Acre        65
25                Amapá        50
26              Roraima        43
```

[ ]:
```
dados_2021.UFOCOR.value_counts(normalize=True)[:5].sum() * 100
```

[ ]: 51.188697187882084

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'UF', suicidio_por_uf_21.head(10),
                  'Número de suicídio por UF - 2021', 'Número de vítimas',␣
 ↪'UF')
```



[ ]:
```
suicidio_por_uf_22 = gerar_dataframe_categorico(dados_2022.UFOCOR, 'UF',␣
 ↪'contagem')
suicidio_por_uf_22
```

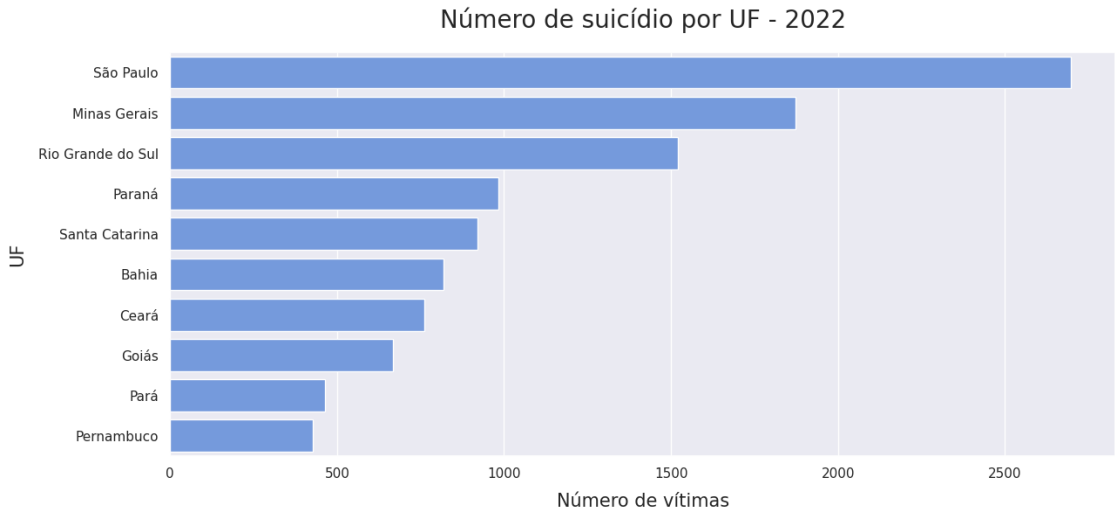
[ ]:

	UF	contagem
0	São Paulo	2696
1	Minas Gerais	1873
2	Rio Grande do Sul	1520
3	Paraná	984
4	Santa Catarina	921
5	Bahia	818
6	Ceará	760
7	Goiás	668
8	Pará	464
9	Pernambuco	427
10	Rio de Janeiro	382
11	Piauí	350
12	Maranhão	341
13	Paraíba	312
14	Mato Grosso do Sul	307
15	Amazonas	291
16	Mato Grosso	286
17	Espírito Santo	224
18	Rio Grande do Norte	221
19	Distrito Federal	208
20	Alagoas	160
21	Sergipe	158
22	Tocantins	146
23	Rondônia	143
24	Acre	71
25	Amapá	59
26	Roraima	40

[ ]:
```
dados_2022.UFOCOR.value_counts(normalize=True)[:5].sum() * 100
```

[ ]: 53.904248145650705

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'UF', suicidio_por_uf_22.head(10),
                  'Número de suicídio por UF - 2022', 'Número de vítimas',␣
 ↪'UF')
```



Em Números absolutos, a distribuição dos estados são razoavelmente estáves na posição. Um destaque é o Estado do Rio de Janeiro que somando os anos de 2019 a 2022 é o oitavo, mas sua posição varia ano a ano, ficando fora dos 10 estados com maiores casos em 2022.

Considerando os números absoutos as escolhas de políticas públicas de saúde serão nestes estados listados, porém uma análise é a quantidade de suicídios em comparação com a população residente naquele estado. Gerando outro enfoque de tratamento.

Para efeitos de população, usaremos a população de 2022 informada pelo IBGE para ter uma proporção relativa à população.

[ ]:
```
populacao_ibge = pd.DataFrame([
              ('São Paulo', 44420459),
              ('Minas Gerais', 20538718),
              ('Rio de Janeiro', 16054524),
              ('Bahia', 14136417),
              ('Paraná', 11443208),
              ('Rio Grande do Sul', 10880506),
              ('Pernambuco', 9058155),
              ('Ceará', 8791688),
              ('Pará', 8116132),
              ('Santa Catarina', 7609601),
              ('Goiás', 7055228),
              ('Maranhão', 6775152),
              ('Paraíba', 3974495),
              ('Amazonas', 3941175),
              ('Espírito Santo', 3833486),
              ('Mato Grosso', 3658813),
              ('Rio Grande do Norte', 3302406),
              ('Piauí', 3269200),
              ('Alagoas', 3127511),
```

```
                ('Distrito Federal', 2817068),
                ('Mato Grosso do Sul', 2756700),
                ('Sergipe', 2209558),
                ('Rondônia', 1581016),
                ('Tocantins', 1511459),
                ('Acre', 830026),
                ('Amapá', 733508),
                ('Roraima', 636303),], columns=['Unidade federativa', 'População␣
 ↪2022'])
populacao_ibge
```

[ ]:

	Unidade federativa	População 2022
0	São Paulo	44420459
1	Minas Gerais	20538718
2	Rio de Janeiro	16054524
3	Bahia	14136417
4	Paraná	11443208
5	Rio Grande do Sul	10880506
6	Pernambuco	9058155
7	Ceará	8791688
8	Pará	8116132
9	Santa Catarina	7609601
10	Goiás	7055228
11	Maranhão	6775152
12	Paraíba	3974495
13	Amazonas	3941175
14	Espírito Santo	3833486
15	Mato Grosso	3658813
16	Rio Grande do Norte	3302406
17	Piauí	3269200
18	Alagoas	3127511
19	Distrito Federal	2817068
20	Mato Grosso do Sul	2756700
21	Sergipe	2209558
22	Rondônia	1581016
23	Tocantins	1511459
24	Acre	830026
25	Amapá	733508
26	Roraima	636303

[ ]:

```
suicidio_por_uf_22_ibge = suicidio_por_uf_22.set_index('UF').
 ↪join(populacao_ibge.set_index('Unidade federativa'))
suicidio_por_uf_22_ibge['suicídio por cem mil habitantes'] =␣
 ↪(suicidio_por_uf_22_ibge['contagem'] / suicidio_por_uf_22_ibge['População␣
 ↪2022']) * 100000
suicidio_por_uf_22_ibge
```

[ ]:

UF	contagem	População 2022	suicídio por cem mil habitantes
São Paulo	2696	44420459	6.069275
Minas Gerais	1873	20538718	9.119362
Rio Grande do Sul	1520	10880506	13.969939
Paraná	984	11443208	8.598987
Santa Catarina	921	7609601	12.103131
Bahia	818	14136417	5.786473
Ceará	760	8791688	8.644529
Goiás	668	7055228	9.468156
Pará	464	8116132	5.717009
Pernambuco	427	9058155	4.713984
Rio de Janeiro	382	16054524	2.379392
Piauí	350	3269200	10.705983
Maranhão	341	6775152	5.033097
Paraíba	312	3974495	7.850054
Mato Grosso do Sul	307	2756700	11.136504
Amazonas	291	3941175	7.383585
Mato Grosso	286	3658813	7.816743
Espírito Santo	224	3833486	5.843246
Rio Grande do Norte	221	3302406	6.692091
Distrito Federal	208	2817068	7.383563
Alagoas	160	3127511	5.115889
Sergipe	158	2209558	7.150751
Tocantins	146	1511459	9.659541
Rondônia	143	1581016	9.044817
Acre	71	830026	8.553949
Amapá	59	733508	8.043539
Roraima	40	636303	6.286313

[ ]:

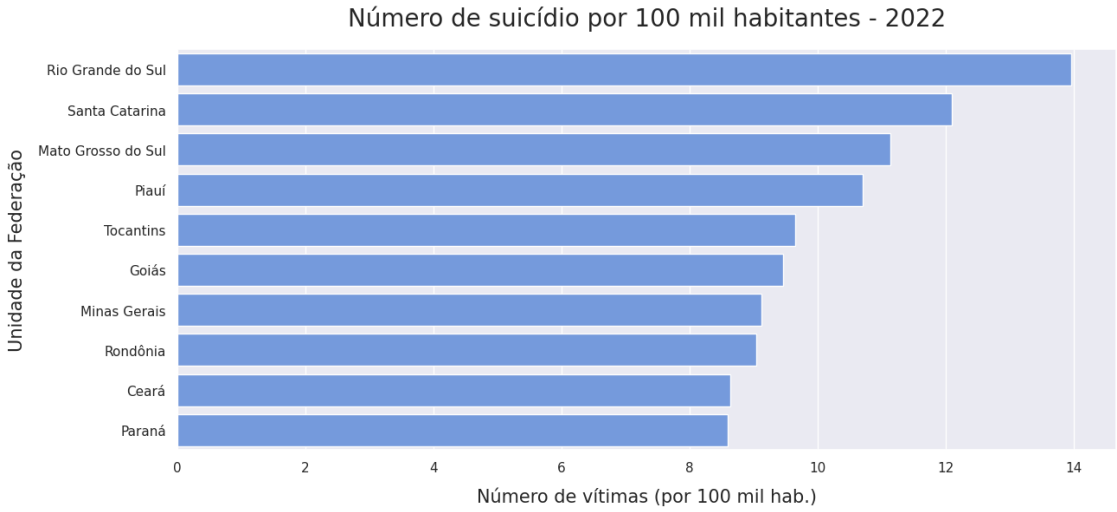
```
suicidio_por_uf_22_ibge.sort_values('suicídio por cem mil habitantes',␣
↪ascending=False)
```

[ ]:

UF	contagem	População 2022	suicídio por cem mil habitantes
Rio Grande do Sul	1520	10880506	13.969939
Santa Catarina	921	7609601	12.103131
Mato Grosso do Sul	307	2756700	11.136504
Piauí	350	3269200	10.705983
Tocantins	146	1511459	9.659541
Goiás	668	7055228	9.468156
Minas Gerais	1873	20538718	9.119362
Rondônia	143	1581016	9.044817
Ceará	760	8791688	8.644529
Paraná	984	11443208	8.598987
Acre	71	830026	8.553949
Amapá	59	733508	8.043539

Paraíba	312	3974495	7.850054
Mato Grosso	286	3658813	7.816743
Amazonas	291	3941175	7.383585
Distrito Federal	208	2817068	7.383563
Sergipe	158	2209558	7.150751
Rio Grande do Norte	221	3302406	6.692091
Roraima	40	636303	6.286313
São Paulo	2696	44420459	6.069275
Espírito Santo	224	3833486	5.843246
Bahia	818	14136417	5.786473
Pará	464	8116132	5.717009
Alagoas	160	3127511	5.115889
Maranhão	341	6775152	5.033097
Pernambuco	427	9058155	4.713984
Rio de Janeiro	382	16054524	2.379392

[ ]:
```
#gráfico de colunas conforme função personalizada

grafico_de_barras('suicídio por cem mil habitantes', 'UF',␣
↪suicidio_por_uf_22_ibge.sort_values('suicídio por cem mil habitantes',␣
↪ascending=False).reset_index().head(10),
                  'Número de suicídio por 100 mil habitantes - 2022', 'Número␣
↪de vítimas (por 100 mil hab.)', 'Unidade da Federação')
```



Se considerarmos o número de habitantes de cada estado, a proporção de casos para cada 100.000 podemos perceber que o Rio de Janeiro é estado com a menor proporção; * 3 estados do sul estão entre os 10 maiores proporções, inclusive os 2 primeiros são da Região Sul do país; * 2 dos 3 estados da Região Centro-Oeste também estão entre os 10 maiores proporções;

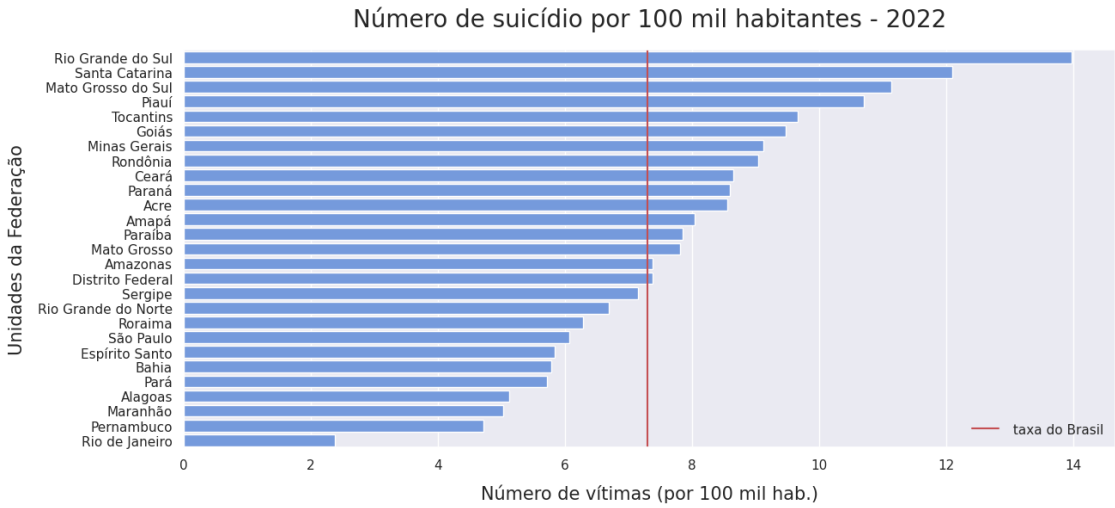
O destaque negativo fica para o Estado do Rio Grande do Sul que é o terceiro estado com o maior

número de suicídios em números absolutos e o primeiro a cada 100 mil habitantes.

[ ]:
```
# Taxa brasileira
taxa_brasil = (suicidio_por_uf_22_ibge.contagem.sum() /␣
↪suicidio_por_uf_22_ibge['População 2022'].sum() ) * 100000
taxa_brasil
```

[ ]: 7.3031697746357045

[ ]:
```
plt.figure(figsize = (14, 6))

plot = sns.barplot(x= 'suicídio por cem mil habitantes', y= 'UF', data=␣
↪suicidio_por_uf_22_ibge.sort_values('suicídio por cem mil habitantes',␣
↪ascending=False).reset_index(), color='#6495ED', orient='h')
plt.title("Número de suicídio por 100 mil habitantes - 2022", fontsize=20,␣
↪pad=20)
plt.xlabel("Número de vítimas (por 100 mil hab.)", fontsize=15, labelpad=10)
plt.ylabel('Unidades da Federação', fontsize=15, labelpad=10)
plt.axvline(x = taxa_brasil, color = 'r', label = ' taxa do Brasil')
plt.legend()
plt.show()
```



2.7 Suicídio por ocupação

[ ]:
```
suicidio_por_ocupacao = gerar_dataframe_categorico(dados_2022.TITULO_OCUPACAO,␣
↪'ocupacao', 'contagem')
suicidio_por_ocupacao
```

[ ]:
```
                                                ocupacao  contagem
0                                               ignorado      6088
1                     Trabalhador agropecuário em geral       773
2                                               Pedreiro       596
3                   Trabalhador volante da agricultura       595
4                     Representante comercial autônomo       537
..                                                    …        …
656  Professor de ensino superior na área de prátic…         1
657      Supervisor de entrevistadores e recenseadores         1
658                              Relojoeiro (reparação)         1
659  Mecânico de manutenção de sistema hidráulico d…         1
660     Analista de crédito (instituições financeiras)         1

[661 rows x 2 columns]
```

[ ]:
```
dados_2022.TITULO_OCUPACAO.value_counts(normalize=True)
```

[ ]:
```
ignorado
0.410519
Trabalhador agropecuário em geral
0.052124
Pedreiro
0.040189
Trabalhador volante da agricultura
0.040121
Representante comercial autônomo
0.036210
            …
Professor de ensino superior na área de prática de ensino
0.000067
Supervisor de entrevistadores e recenseadores
0.000067
Relojoeiro (reparação)
0.000067
Mecânico de manutenção de sistema hidráulico de aeronaves (serviços de pista e
hangar)    0.000067
Analista de crédito (instituições financeiras)
0.000067
Name: TITULO_OCUPACAO, Length: 661, dtype: float64
```

Como essa informação é importante para a análise, mas tem muitas observações faltantes : ignorado, faremos uma análise retirando esse valor

[ ]:
```
suicidio_por_ocupacao = pd.DataFrame(dados.TITULO_OCUPACAO.value_counts()[1:])
suicidio_por_ocupacao = suicidio_por_ocupacao.reset_index()
suicidio_por_ocupacao.columns = ['ocupacao', 'contagem']
suicidio_por_ocupacao
```

[ ]:
```
                                          ocupacao  contagem
0                    Trabalhador agropecuário em geral      2896
1                   Trabalhador volante da agricultura      2512
2                                             Pedreiro      2047
3                     Representante comercial autônomo      1837
4             Empregado  doméstico  nos serviços gerais      1334
…                                                   …       …
1078  Trabalhador na olericultura (raízes, bulbos e …         1
1079                             Gerente de comunicação         1
1080                              Psicólogo educacional         1
1081                      Chapeleiro (chapéus de palha)         1
1082                      Diretor de arte (publicidade)         1

[1083 rows x 2 columns]
```

[ ]:
```
suicidio_por_ocupacao['percentual'] = (suicidio_por_ocupacao.contagem /␣
↪suicidio_por_ocupacao.contagem.sum()) * 100
suicidio_por_ocupacao
```

[ ]:
```
                                          ocupacao  contagem  percentual
0                    Trabalhador agropecuário em geral      2896    8.976783
1                   Trabalhador volante da agricultura      2512    7.786491
2                                             Pedreiro      2047    6.345123
3                     Representante comercial autônomo      1837    5.694182
4             Empregado  doméstico  nos serviços gerais      1334    4.135024
…                                                   …       …          …
1078  Trabalhador na olericultura (raízes, bulbos e …         1    0.003100
1079                             Gerente de comunicação         1    0.003100
1080                              Psicólogo educacional         1    0.003100
1081                      Chapeleiro (chapéus de palha)         1    0.003100
1082                      Diretor de arte (publicidade)         1    0.003100

[1083 rows x 3 columns]
```

2.7.1 20 Ocupações com maior incidência de suicídio anos 2019 a 2022

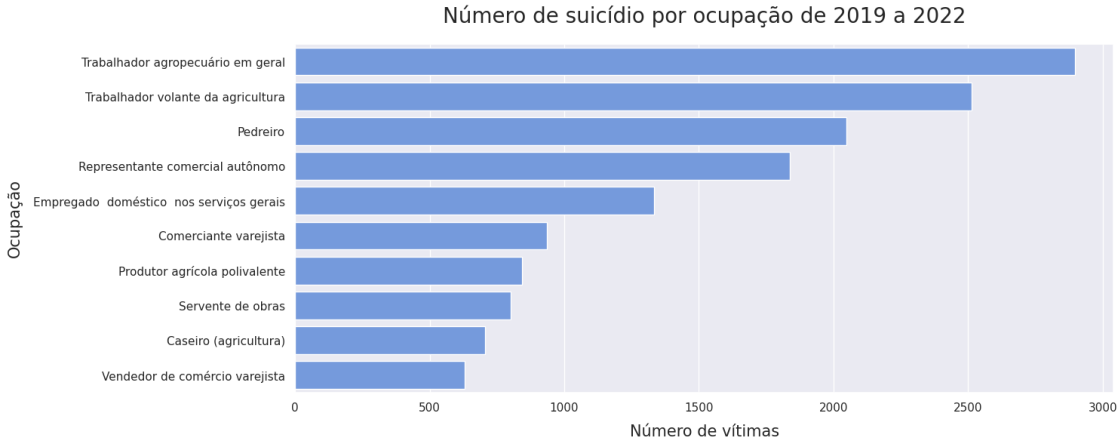
[ ]:
```
suicidio_por_ocupacao.head(20)
```

[ ]:
```
                                          ocupacao  contagem  percentual
0                    Trabalhador agropecuário em geral      2896    8.976783
1                   Trabalhador volante da agricultura      2512    7.786491
2                                             Pedreiro      2047    6.345123
3                     Representante comercial autônomo      1837    5.694182
4             Empregado  doméstico  nos serviços gerais      1334    4.135024
5                                Comerciante varejista       935    2.898236
6                       Produtor agrícola polivalente       843    2.613062
7                                   Servente de obras       802    2.485974
```

```
8                                Caseiro (agricultura)       705    2.185301
9                       Vendedor de comércio varejista       631    1.955922
10                       Motorista de carro de passeio       575    1.782338
11                                    Pintor de obras       419    1.298782
12  Mecânico de manutenção de automóveis, motocicl…       376    1.165494
13                                          Vigilante       374    1.159295
14  Motorista de caminhão (rotas regionais e inter…       356    1.103500
15                    Alimentador de linha de produção       308    0.954713
16                                           Advogado       295    0.914417
17                          Assistente administrativo       280    0.867921
18                              Técnico de enfermagem       273    0.846223
19                         Eletricista de instalações       242    0.750132
```

Considerando as 20 ocupações com maior número de pessoas que se suicidaram, podemos detacar os seguintes pontos: * A profissão de advogado é a única de nível superior; * Há a presença grande de trabalhadores do meio rural, inclusive os dois maiores; * Presença significativa de ocupações ligadoas À construção civil * Com a exceção de advogado, as ocupações são de perfil de baixa escolaridade. o que corrobora o achado anterior na análise de escolaridade. * empregados domésticos em 5º também é de se pontuar

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'ocupacao', suicidio_por_ocupacao.head(10),
                  'Número de suicídio por ocupação de 2019 a 2022', 'Número de␣
 ↪vítimas', 'Ocupação')
```

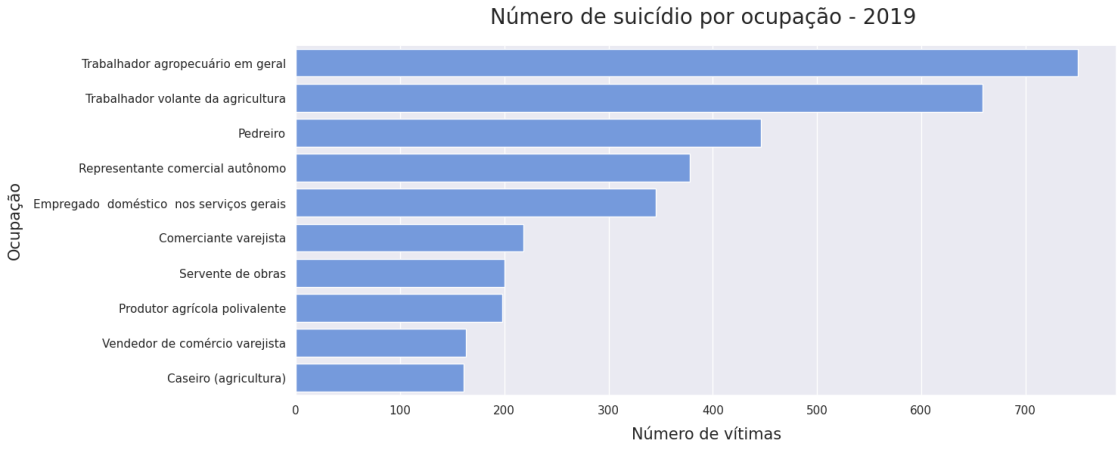


[ ]:
```
#2019
suicidio_por_ocupacao_19 = pd.DataFrame(dados_2019.TITULO_OCUPACAO.
 ↪value_counts()[1:])
suicidio_por_ocupacao_19 = suicidio_por_ocupacao_19.reset_index()
suicidio_por_ocupacao_19.columns = ['ocupacao', 'contagem']
suicidio_por_ocupacao_19.head(20)
```

[ ]:

	ocupacao	contagem
0	Trabalhador agropecuário em geral	750
1	Trabalhador volante da agricultura	659
2	Pedreiro	446
3	Representante comercial autônomo	378
4	Empregado doméstico nos serviços gerais	345
5	Comerciante varejista	218
6	Servente de obras	200
7	Produtor agrícola polivalente	198
8	Vendedor de comércio varejista	163
9	Caseiro (agricultura)	161
10	Motorista de carro de passeio	122
11	Mecânico de manutenção de automóveis, motocicl…	110
12	Pintor de obras	105
13	Vigilante	85
14	Motorista de caminhão (rotas regionais e inter…	83
15	Advogado	73
16	Alimentador de linha de produção	69
17	Eletricista de instalações	59
18	Produtor agropecuário, em geral	57
19	Técnico de enfermagem	55

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'ocupacao', suicidio_por_ocupacao_19.head(10),
                  'Número de suicídio por ocupação - 2019 ', 'Número de␣
 ↪vítimas', 'Ocupação')
```



[ ]:
```
#2020
suicidio_por_ocupacao_20 = pd.DataFrame(dados_2020.TITULO_OCUPACAO.
 ↪value_counts()[1:])
suicidio_por_ocupacao_20 = suicidio_por_ocupacao_20.reset_index()
```

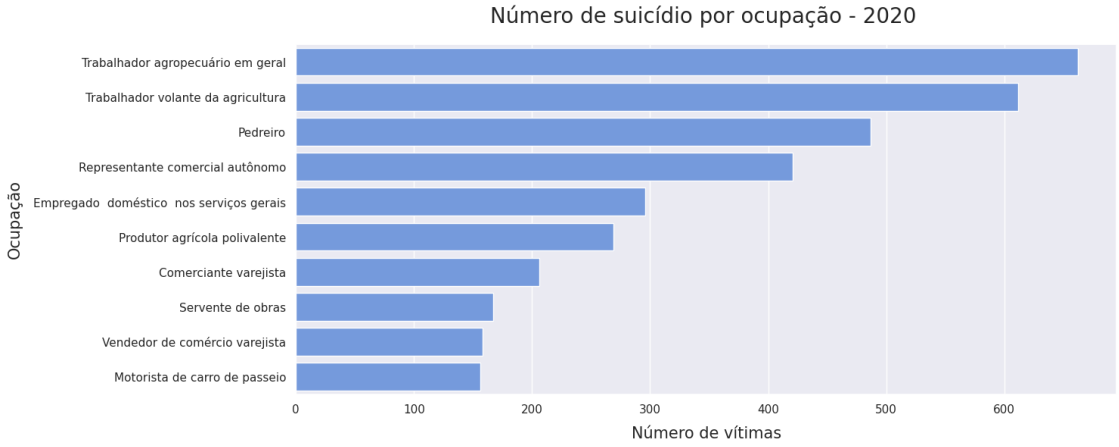
```
suicidio_por_ocupacao_20.columns = ['ocupacao', 'contagem']
suicidio_por_ocupacao_20.head(20)
```

[ ]:

	ocupacao	contagem
0	Trabalhador agropecuário em geral	662
1	Trabalhador volante da agricultura	612
2	Pedreiro	487
3	Representante comercial autônomo	421
4	Empregado doméstico nos serviços gerais	296
5	Produtor agrícola polivalente	269
6	Comerciante varejista	206
7	Servente de obras	167
8	Vendedor de comércio varejista	158
9	Motorista de carro de passeio	156
10	Caseiro (agricultura)	148
11	Pintor de obras	92
12	Motorista de caminhão (rotas regionais e inter…	85
13	Vigilante	81
14	Mecânico de manutenção de automóveis, motocicl…	78
15	Advogado	75
16	Alimentador de linha de produção	69
17	Técnico de enfermagem	68
18	Cabeleireiro	66
19	Assistente administrativo	66

[ ]:

```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'ocupacao', suicidio_por_ocupacao_20.head(10),
                  'Número de suicídio por ocupação - 2020 ', 'Número de␣
 ↪vítimas', 'Ocupação')
```

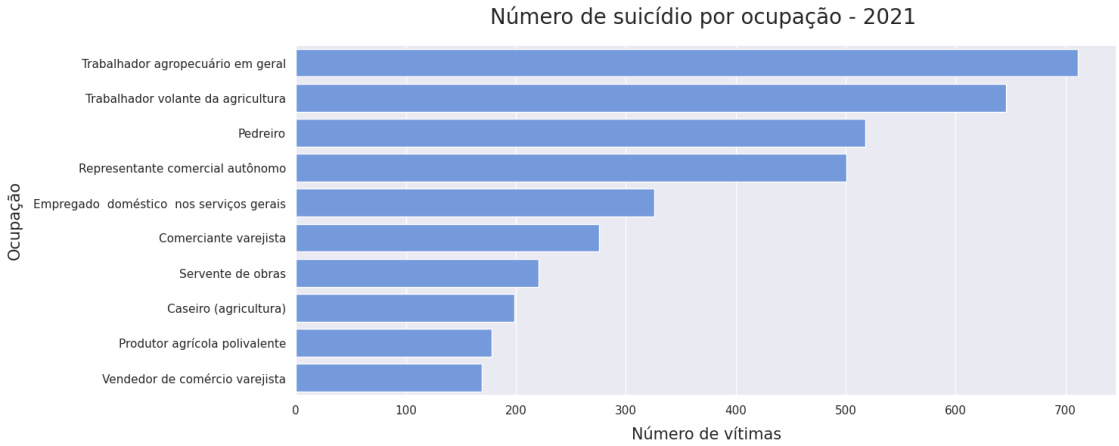


[ ]:
```
#2021
suicidio_por_ocupacao_21 = pd.DataFrame(dados_2021.TITULO_OCUPACAO.
 ↪value_counts()[1:])
suicidio_por_ocupacao_21 = suicidio_por_ocupacao_21.reset_index()
suicidio_por_ocupacao_21.columns = ['ocupacao', 'contagem']
suicidio_por_ocupacao_21.head(20)
```

[ ]:

	ocupacao	contagem
0	Trabalhador agropecuário em geral	711
1	Trabalhador volante da agricultura	646
2	Pedreiro	518
3	Representante comercial autônomo	501
4	Empregado doméstico nos serviços gerais	326
5	Comerciante varejista	276
6	Servente de obras	221
7	Caseiro (agricultura)	199
8	Produtor agrícola polivalente	178
9	Vendedor de comércio varejista	169
10	Motorista de carro de passeio	141
11	Vigilante	111
12	Pintor de obras	104
13	Mecânico de manutenção de automóveis, motocicl…	94
14	Alimentador de linha de produção	82
15	Assistente administrativo	80
16	Advogado	80
17	Motorista de caminhão (rotas regionais e inter…	77
18	Técnico de enfermagem	73
19	Operador de máquinas fixas, em geral	70

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'ocupacao', suicidio_por_ocupacao_21.head(10),
                  'Número de suicídio por ocupação - 2021 ', 'Número de␣
 ↪vítimas', 'Ocupação')
```

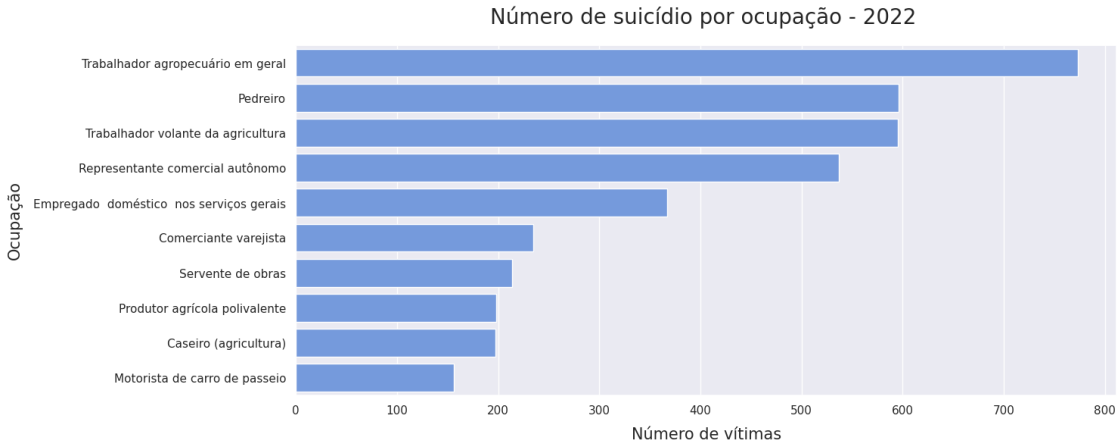


[ ]:
```
#2022
suicidio_por_ocupacao_22 = pd.DataFrame(dados_2022.TITULO_OCUPACAO.
 ↪value_counts()[1:])
suicidio_por_ocupacao_22 = suicidio_por_ocupacao_22.reset_index()
suicidio_por_ocupacao_22.columns = ['ocupacao', 'contagem']
suicidio_por_ocupacao_22.head(20)
```

[ ]:

	ocupacao	contagem
0	Trabalhador agropecuário em geral	773
1	Pedreiro	596
2	Trabalhador volante da agricultura	595
3	Representante comercial autônomo	537
4	Empregado doméstico nos serviços gerais	367
5	Comerciante varejista	235
6	Servente de obras	214
7	Produtor agrícola polivalente	198
8	Caseiro (agricultura)	197
9	Motorista de carro de passeio	156
10	Vendedor de comércio varejista	141
11	Pintor de obras	118
12	Motorista de caminhão (rotas regionais e inter…	111
13	Vigilante	97
14	Mecânico de manutenção de automóveis, motocicl…	94
15	Alimentador de linha de produção	88
16	Assistente administrativo	85
17	Técnico de enfermagem	77
18	Eletricista de instalações	74
19	Operador de máquinas fixas, em geral	72

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'ocupacao', suicidio_por_ocupacao_22.head(10),
                  'Número de suicídio por ocupação - 2022 ', 'Número de␣
 ↪vítimas', 'Ocupação')
```



Somente no ano de 2022 a ocupação de advogado saiu das 20 maiores profissões.

Entretanto, as demais profissões continuam predominantes na listagem desde 2019

2.8 Suicídio por idade

Mesmo havendo uma coluna de idade nos dados originais, estes são codificados conforme regra presente no dicionário e documentação dos dados

[ ]:
```
# manipulação necessária quando usando dados do csv
dados['DTOBITO'] = pd.to_datetime(dados['DTOBITO'])
dados['DTNASC'] = pd.to_datetime(dados['DTNASC'])

dados_2019['DTOBITO'] = pd.to_datetime(dados_2019['DTOBITO'])
dados_2020['DTOBITO'] = pd.to_datetime(dados_2020['DTOBITO'])
dados_2021['DTOBITO'] = pd.to_datetime(dados_2021['DTOBITO'])
dados_2022['DTOBITO'] = pd.to_datetime(dados_2022['DTOBITO'])

dados_2019['DTNASC'] = pd.to_datetime(dados_2019['DTNASC'])
dados_2020['DTNASC'] = pd.to_datetime(dados_2020['DTNASC'])
dados_2021['DTNASC'] = pd.to_datetime(dados_2021['DTNASC'])
dados_2022['DTNASC'] = pd.to_datetime(dados_2022['DTNASC'])
```

[ ]:
```
dados['idade_calculada'] = (dados['DTOBITO'] - dados['DTNASC']) / np.
↪timedelta64(1, 'Y')
dados['idade_calculada'] = dados['idade_calculada'].astype(int)
dados.head()
```

[ ]:
```
   CODUFOCOR CAUSABAS      OCUP    DTOBITO HORAOBITO  NATURAL  CODMUNNATU  \
0         13     X700  612005.0 2019-01-01  18:01:00    813.0    130006.0
1         51     X700  999991.0 2019-01-01  05:01:00    851.0    510395.0
2         28     X700  999991.0 2019-01-01  06:01:00    829.0    292400.0
3         13     X700  999994.0 2019-01-01  02:01:00    813.0    130290.0
```

```
4         16     X702  999991.0 2019-01-01  20:01:00    815.0    150050.0

      DTNASC  IDADE       SEXO   RACACOR    ESTCIV             ESC  \
0 1994-09-07    424  masculino  indigena  solteiro  de 8 a 11 anos
1 2004-09-11    414   feminino    branca  solteiro  de 8 a 11 anos
2 2000-07-03    418  masculino     parda  solteiro   de 4 a 7 anos
3 1985-11-28    433  masculino     parda  solteiro         nenhuma
4 1997-01-24    421   feminino     parda  solteiro  de 8 a 11 anos

                         ESC2010  SERIESCFAL  CODMUNRES    LOCOCOR  \
0          médio (antigo 2º grau)         NaN     130006  domicílio
1          médio (antigo 2º grau)         NaN     510395  domicílio
2  fundamental II (5ª a 8ª série)         5.0     280450  domicílio
3                sem escolaridade         NaN     130290  domicílio
4          médio (antigo 2º grau)         NaN     150050   hospital

   CODMUNOCOR ASSISTMED NECROPSIA  COMUNSVOIM  DTATESTADO CIRCOBITO  \
0      130006       não       não         NaN  2019-01-03  suicídio
1      510395       NaN       NaN    510250.0  2019-01-01  suicídio
2      280450       não       sim    280030.0  2019-01-01  suicídio
3      130290       sim       não         NaN  2019-01-01  suicídio
4      160030       não       sim    160030.0  2019-01-02  suicídio

                FONTE  TPPOS CAUSABAS_O                 ESCFALAGR1  CONTADOR  \
0             família    NaN       X700    ensino médio incompleto    510709
1  ocorrência policial  False       X700    ensino médio incompleto    185795
2  ocorrência policial  False       X700  fundamental II incompleto    733808
3             família    NaN       X700           sem escolaridade    137583
4  ocorrência policial    NaN       X702    ensino médio incompleto    278318

                TITULO_OCUPACAO  \
0  Produtor agrícola polivalente
1                      ignorado
2                      ignorado
3                      ignorado
4                      ignorado

                                            CID10       UFOCOR   ANO  \
0  Lesão autoprovocada intencionalmente por enfor…     Amazonas  2019
1  Lesão autoprovocada intencionalmente por enfor…  Mato Grosso  2019
2  Lesão autoprovocada intencionalmente por enfor…      Sergipe  2019
3  Lesão autoprovocada intencionalmente por enfor…     Amazonas  2019
4  Lesão autoprovocada intencionalmente por enfor…        Amapá  2019

   idade_calculada
0               24
1               14
```

```
2                18
3                33
4                21
```

[ ]:
```python
suicidio_por_idade = gerar_dataframe_categorico(dados.idade_calculada, 'idade',␣
↪'contagem')
suicidio_por_idade
```

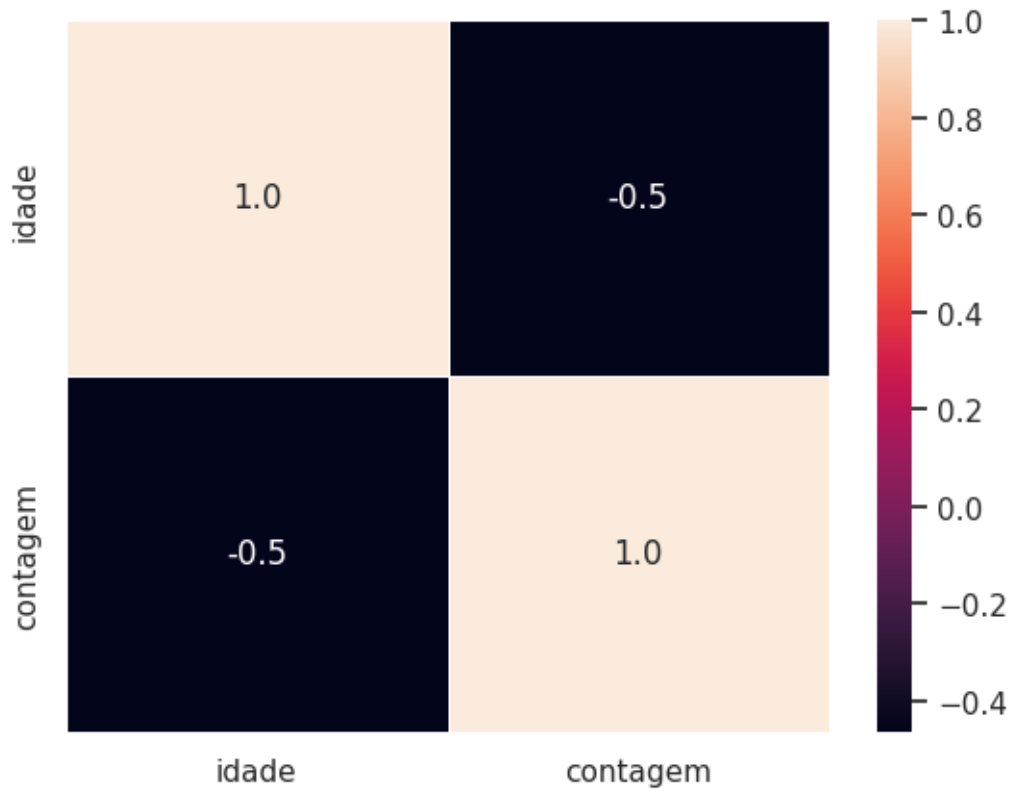
[ ]:
```
    idade  contagem
0      24      1215
1      21      1213
2      37      1211
3      22      1204
4      23      1182
..    …        …
92     99         2
93      5         1
94     98         1
95      0         1
96      6         1

[97 rows x 2 columns]
```

[ ]:
```python
suicidio_por_idade.corr()
```

[ ]:
```
              idade  contagem
idade     1.000000 -0.462772
contagem -0.462772  1.000000
```

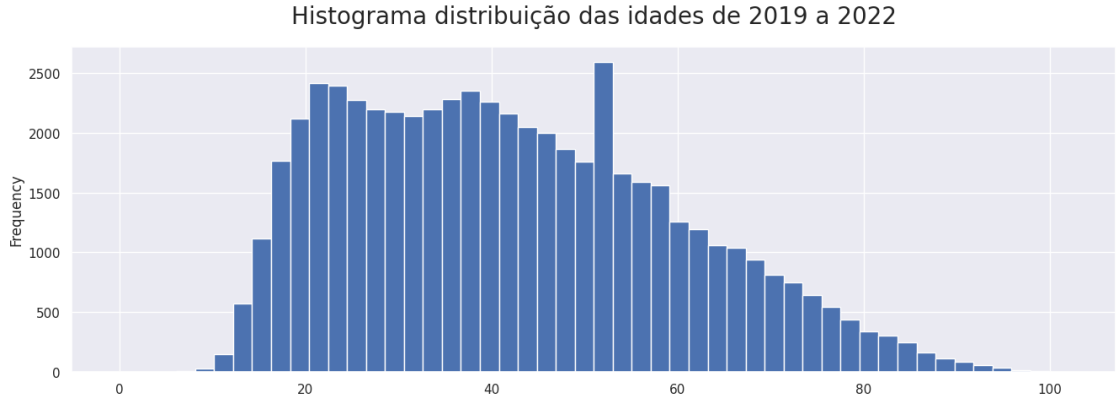
[ ]:
```python
sns.heatmap(suicidio_por_idade.corr(), annot = True, fmt=".1f", linewidths=.6)
```

[ ]:
```
<Axes: >
```



A correlação entre idade e quantidade de suicídios é baixa

[ ]:
```
plt.figure(figsize = (16, 5))

dados.idade_calculada.plot(kind='hist', bins=50, xlabel="Anos")
plt.title(label="Histograma distribuição das idades de 2019 a 2022",␣
 ↪fontsize=20, pad=20)
plt.show()
```



[ ]:
```
suicidio_por_idade.head(20)
```

[ ]:

	idade	contagem
0	24	1215
1	21	1213
2	37	1211
3	22	1204
4	23	1182
5	39	1163
6	35	1150
7	26	1148
8	38	1141
9	36	1137
10	25	1128
11	41	1121
12	27	1116
13	34	1104
14	29	1101
15	40	1097
16	33	1095
17	28	1086
18	32	1080
19	30	1077

[ ]:
```
pd.DataFrame(dados.idade_calculada.describe())
```

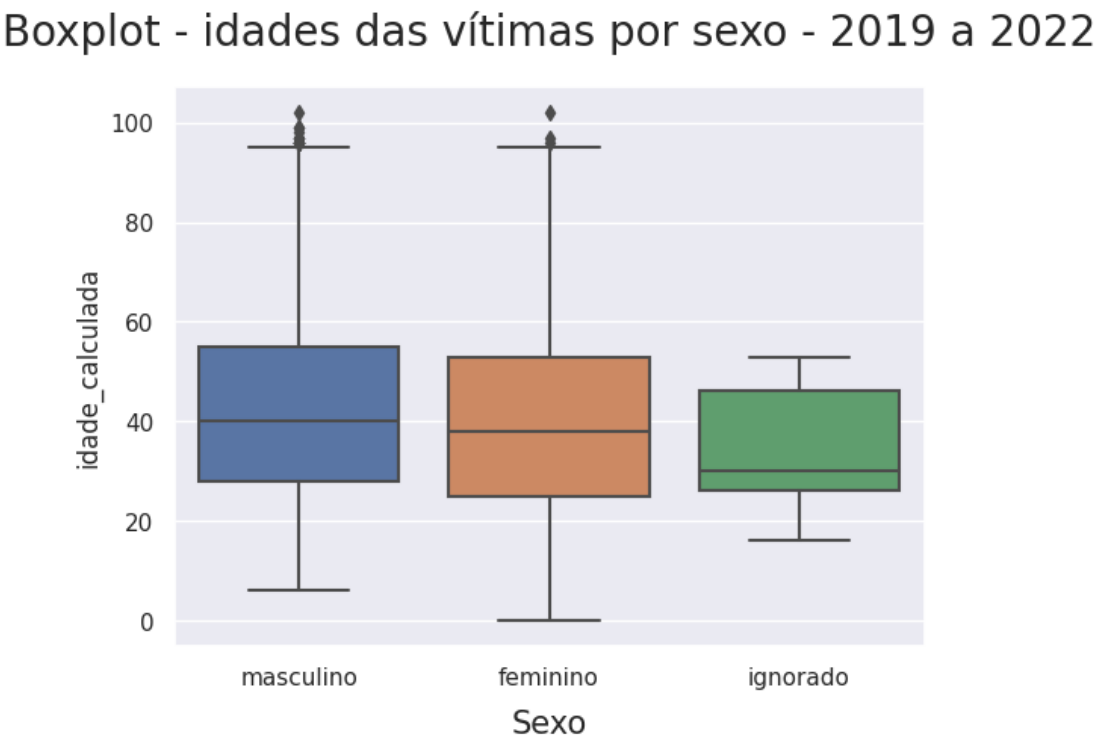
[ ]:

	idade_calculada
count	55727.000000
mean	42.067759
std	17.629341
min	0.000000
25%	27.000000
50%	40.000000
75%	55.000000
max	102.000000

Acredita-se que essa idade de 0 anos seja um erro no preenchimento da tada de nascimento da pessoa. É de se destacar a idade máxima de 102 anos e que 50% dos dados estejam entre 27 e 55 anos.

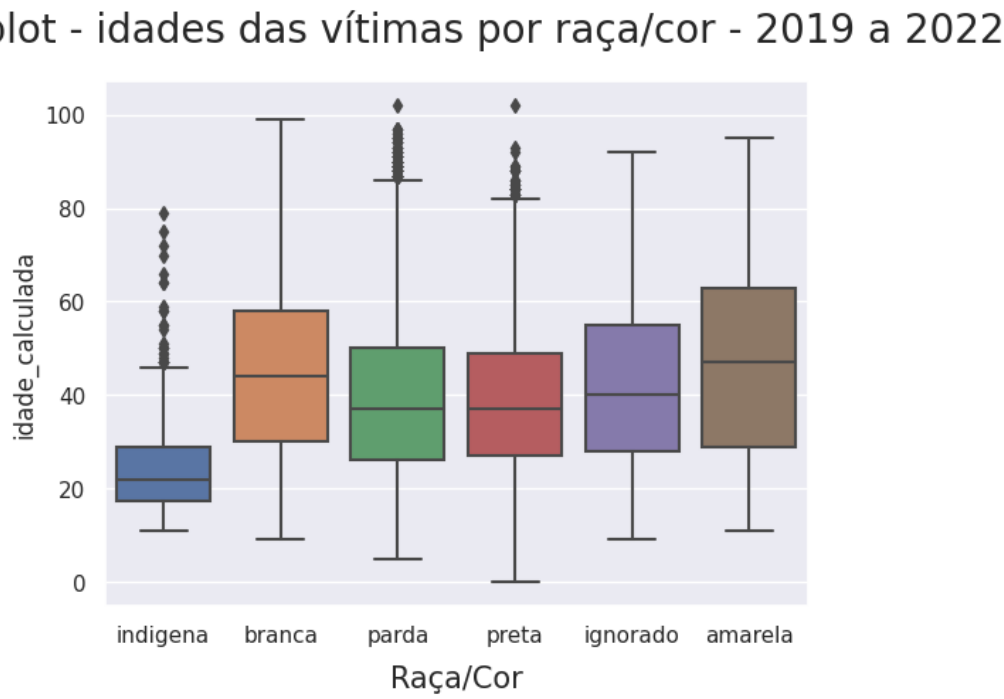
2.8.1 Suicídio por idade e sexo

[ ]:
```
grafico_de_caixa(x='SEXO', y='idade_calculada', dados=dados,
                 nome_grafico="Boxplot - idades das vítimas por sexo - 2019 a␣
 ↪2022",
                rotulo_x="Sexo")
```



Não é possível identificar diferença da distribuição de idades e o sexo

[ ]:
```
grafico_de_caixa(x='RACACOR', y='idade_calculada', dados=dados,
                 nome_grafico="Boxplot - idades das vítimas por raça/cor - 2019␣
 ↪a 2022",
                rotulo_x="Raça/Cor")
```



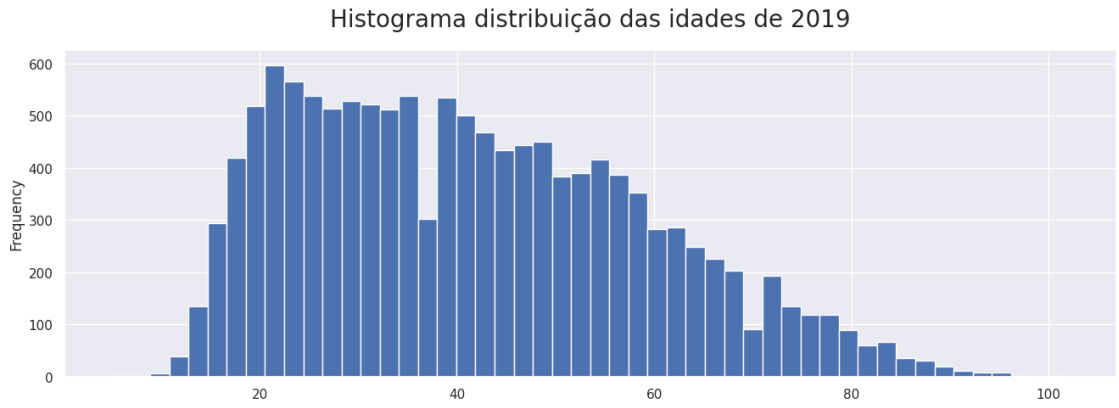
Embora os dados não tenham grande diferença, é perceptível que as pessoas indígenas se suicidam quando são mais novos e os amarelos levemente mais velhos.

[ ]:
```
dados_2019['idade_calculada'] = (dados_2019['DTOBITO'] - dados_2019['DTNASC']) /
 ↪ np.timedelta64(1, 'Y')
dados_2019['idade_calculada'] = dados_2019['idade_calculada'].astype(int)
dados_2019.head()
```

[ ]:
```
   CODUFOCOR CAUSABAS      OCUP    DTOBITO HORAOBITO  NATURAL  CODMUNNATU  \
0         13     X700  612005.0 2019-01-01  18:01:00    813.0    130006.0
1         51     X700  999991.0 2019-01-01  05:01:00    851.0    510395.0
2         28     X700  999991.0 2019-01-01  06:01:00    829.0    292400.0
3         13     X700  999994.0 2019-01-01  02:01:00    813.0    130290.0
4         16     X702  999991.0 2019-01-01  20:01:00    815.0    150050.0

      DTNASC  IDADE       SEXO   RACACOR    ESTCIV             ESC  \
0 1994-09-07    424  masculino  indigena  solteiro  de 8 a 11 anos
1 2004-09-11    414   feminino    branca  solteiro  de 8 a 11 anos
2 2000-07-03    418  masculino     parda  solteiro   de 4 a 7 anos
3 1985-11-28    433  masculino     parda  solteiro         nenhuma
4 1997-01-24    421   feminino     parda  solteiro  de 8 a 11 anos

                          ESC2010  SERIESCFAL  CODMUNRES    LOCOCOR  \
0          médio (antigo 2º grau)         NaN     130006  domicílio
```

```
1          médio (antigo 2º grau)         NaN     510395  domicílio
2  fundamental II (5ª a 8ª série)         5.0     280450  domicílio
3                sem escolaridade         NaN     130290  domicílio
4          médio (antigo 2º grau)         NaN     150050   hospital

   CODMUNOCOR ASSISTMED NECROPSIA  COMUNSVOIM  DTATESTADO CIRCOBITO  \
0      130006       não       não         NaN  2019-01-03  suicídio
1      510395       NaN       NaN    510250.0  2019-01-01  suicídio
2      280450       não       sim    280030.0  2019-01-01  suicídio
3      130290       sim       não         NaN  2019-01-01  suicídio
4      160030       não       sim    160030.0  2019-01-02  suicídio

                FONTE  TPPOS CAUSABAS_O                 ESCFALAGR1  CONTADOR  \
0             família    NaN       X700    ensino médio incompleto    510709
1  ocorrência policial  False       X700    ensino médio incompleto    185795
2  ocorrência policial  False       X700  fundamental II incompleto    733808
3             família    NaN       X700           sem escolaridade    137583
4  ocorrência policial    NaN       X702    ensino médio incompleto    278318

                TITULO_OCUPACAO  \
0  Produtor agrícola polivalente
1                       ignorado
2                       ignorado
3                       ignorado
4                       ignorado

                                            CID10       UFOCOR   ANO  \
0  Lesão autoprovocada intencionalmente por enfor…     Amazonas  2019
1  Lesão autoprovocada intencionalmente por enfor…  Mato Grosso  2019
2  Lesão autoprovocada intencionalmente por enfor…      Sergipe  2019
3  Lesão autoprovocada intencionalmente por enfor…     Amazonas  2019
4  Lesão autoprovocada intencionalmente por enfor…        Amapá  2019

   idade_calculada
0               24
1               14
2               18
3               33
4               21
```

[ ]:
```
plt.figure(figsize = (16, 5))

dados_2019.idade_calculada.plot(kind='hist', bins=50, xlabel="Anos")
plt.title(label="Histograma distribuição das idades de 2019", fontsize=20,␣
 ↪pad=20)
plt.show()
```



Anomalia na distribuição entre 38 a 40 anos

[ ]:
```
pd.DataFrame(dados_2019.idade_calculada.describe())
```

[ ]:
```
       idade_calculada
count     13009.000000
mean         41.537397
std          17.494315
min           5.000000
25%          27.000000
50%          39.000000
75%          54.000000
max         102.000000
```

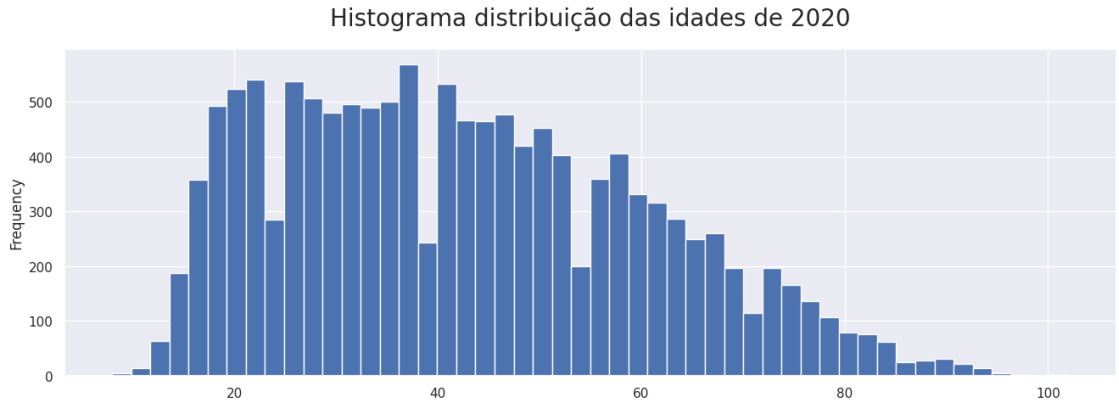
[ ]:
```
dados_2020['idade_calculada'] = (dados_2020['DTOBITO'] - dados_2020['DTNASC']) /
 ↪ np.timedelta64(1, 'Y')
dados_2020['idade_calculada'] = dados_2020['idade_calculada'].astype(int)
dados_2020.head()
```

[ ]:
```
       CODUFOCOR CAUSABAS      OCUP    DTOBITO HORAOBITO  NATURAL  CODMUNNATU  \
13009         35     X680  621005.0 2020-01-01  16:01:00    835.0    355470.0
13010         31     X800  999991.0 2020-01-01  18:01:00    831.0    310350.0
13011         43     X700  724415.0 2020-01-01  00:01:00    843.0    432250.0
13012         31     X709  782410.0 2020-01-01  01:01:00      NaN         NaN
13013         41     X700  784205.0 2020-01-01  07:01:00    841.0    410830.0

          DTNASC  IDADE       SEXO   RACACOR    ESTCIV             ESC  \
13009 1984-07-15    435  masculino    branca    casado  de 8 a 11 anos
13010 2000-10-11    419  masculino    branca  solteiro  12 anos e mais
13011 1984-04-30    435  masculino    branca  solteiro        ignorado
13012 1991-08-06    428  masculino  ignorado  solteiro  de 8 a 11 anos
13013 1999-04-02    420  masculino     preta  solteiro   de 4 a 7 anos
```

```
                            ESC2010  SERIESCFAL  CODMUNRES      LOCOCOR  \
13009  fundamental II (5ª a 8ª série)         8.0     352590    domicílio
13010             superior incompleto         NaN     310350  via pública
13011                        ignorado         NaN     432250    domicílio
13012          médio (antigo 2º grau)         3.0     315733       outros
13013  fundamental II (5ª a 8ª série)         NaN     412545       outros

       CODMUNOCOR ASSISTMED NECROPSIA  COMUNSVOIM  DTATESTADO CIRCOBITO  \
13009      352590       não       sim    352590.0  2020-01-02  suicídio
13010      310350       não       sim    310350.0  2020-01-01  suicídio
13011      432250       NaN       sim    432250.0  2020-01-03  suicídio
13012      316250       NaN       NaN    316250.0  2020-01-02  suicídio
13013      410753       não       sim    412770.0  2020-01-01  suicídio

                    FONTE  TPPOS CAUSABAS_O                 ESCFALAGR1  \
13009  ocorrência policial  False       X680    fundamental II completo
13010                outra    NaN       X800           sem escolaridade
13011  ocorrência policial    NaN       X700                   ignorado
13012             hospital    NaN       X709      ensino médio completo
13013  ocorrência policial  False       X700  fundamental II incompleto

       CONTADOR                    TITULO_OCUPACAO  \
13009    437850  Trabalhador agropecuário em geral
13010    433576                           ignorado
13011    430773                          Chapeador
13012    227308         Motorista de ônibus urbano
13013    290672   Alimentador de linha de produção

                                                CID10             UFOCOR  \
13009  Auto-intoxicação por e exposição, intencional,…          São Paulo
13010  Lesão autoprovocada intencionalmente por preci…       Minas Gerais
13011  Lesão autoprovocada intencionalmente por enfor…  Rio Grande do Sul
13012  Lesão autoprovocada intencionalmente por enfor…       Minas Gerais
13013  Lesão autoprovocada intencionalmente por enfor…             Paraná

        ANO  idade_calculada
13009  2020               35
13010  2020               19
13011  2020               35
13012  2020               28
13013  2020               20
```

[ ]:
```
plt.figure(figsize = (16, 5))

dados_2020.idade_calculada.plot(kind='hist', bins=50, xlabel="Anos")
plt.title(label="Histograma distribuição das idades de 2020", fontsize=20,␣
 ↪pad=20)
```

```
plt.show()
```



[ ]:
```
pd.DataFrame(dados_2020.idade_calculada.describe())
```

[ ]:
```
       idade_calculada
count     13166.000000
mean         42.589701
std          17.878090
min           8.000000
25%          28.000000
50%          41.000000
75%          55.000000
max         102.000000
```

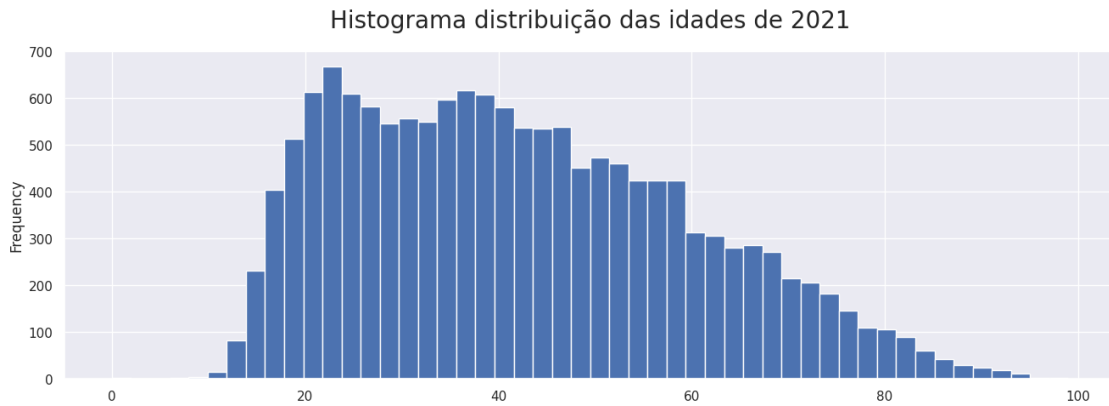
[ ]:
```
dados_2021['idade_calculada'] = (dados_2021['DTOBITO'] - dados_2021['DTNASC']) /
 ↪ np.timedelta64(1, 'Y')
dados_2021['idade_calculada'] = dados_2021['idade_calculada'].astype(int)
dados_2021.head()
```

[ ]:
```
       CODUFOCOR CAUSABAS      OCUP    DTOBITO HORAOBITO  NATURAL  CODMUNNATU  \
26175         43     X700  783210.0 2021-01-01  23:01:00    843.0    431490.0
26176         25     X700       NaN 2021-01-01  14:01:00    825.0    250700.0
26177         51     X700  725010.0 2021-01-01  00:01:00    851.0    510340.0
26178         43     X700  716610.0 2021-01-01  02:01:00    843.0    431490.0
26179         41     X708  782305.0 2021-01-01  16:01:00    841.0    410690.0

          DTNASC  IDADE       SEXO RACACOR    ESTCIV             ESC  \
26175 1993-01-05    427  masculino  branca  solteiro  de 8 a 11 anos
26176 1964-03-03    456  masculino   parda  ignorado        ignorado
26177 1986-11-24    434  masculino   parda  solteiro   de 4 a 7 anos
26178 1976-09-26    444  masculino  branca    casado   de 1 a 3 anos
26179 1983-12-11    437  masculino  branca  solteiro  de 8 a 11 anos
```

```
                              ESC2010  SERIESCFAL  CODMUNRES    LOCOCOR  \
26175          médio (antigo 2º grau)         NaN     430920  domicílio
26176                        ignorado         NaN     250700  domicílio
26177  fundamental II (5ª a 8ª série)         NaN     510840  domicílio
26178   fundamental I (1ª a 4ª série)         NaN     430460  domicílio
26179  fundamental II (5ª a 8ª série)         8.0     410690     outros

       CODMUNOCOR ASSISTMED NECROPSIA  COMUNSVOIM  DTATESTADO CIRCOBITO  \
26175      430920  ignorado       sim    431490.0  2021-01-02  suicídio
26176      250700       NaN       NaN    251080.0  2021-01-02  suicídio
26177      510840       não       sim    510340.0  2021-01-05  suicídio
26178      430460       não       sim    431490.0  2021-01-01  suicídio
26179      410690       não       sim    410690.0  2021-01-01  suicídio

                    FONTE  TPPOS CAUSABAS_O                 ESCFALAGR1  \
26175  ocorrência policial  False       X700    ensino médio incompleto
26176  ocorrência policial    NaN       X700                   ignorado
26177  ocorrência policial    NaN       X700  fundamental II incompleto
26178               outra    NaN       Y200   fundamental I incompleto
26179             ignorado  False       X708    fundamental II completo

       CONTADOR                TITULO_OCUPACAO  \
26175    303272           Carregador (armazém)
26176    982235                       ignorado
26177    317052             Ajustador mecânico
26178    289254                Pintor de obras
26179    679207  Motorista de carro de passeio

                                                CID10             UFOCOR  \
26175  Lesão autoprovocada intencionalmente por enfor…  Rio Grande do Sul
26176  Lesão autoprovocada intencionalmente por enfor…            Paraíba
26177  Lesão autoprovocada intencionalmente por enfor…        Mato Grosso
26178  Lesão autoprovocada intencionalmente por enfor…  Rio Grande do Sul
26179  Lesão autoprovocada intencionalmente por enfor…             Paraná

        ANO  idade_calculada
26175  2021               27
26176  2021               56
26177  2021               34
26178  2021               44
26179  2021               37
```

[ ]:
```
plt.figure(figsize = (16, 5))

dados_2021.idade_calculada.plot(kind='hist', bins=50, xlabel="Anos")
```

```
plt.title(label="Histograma distribuição das idades de 2021", fontsize=20,␣
↪pad=20)
plt.show()
```



[ ]:
```
pd.DataFrame(dados_2021.idade_calculada.describe())
```

[ ]:
```
       idade_calculada
count     14722.000000
mean         42.187814
std          17.773128
min           0.000000
25%          27.000000
50%          40.000000
75%          55.000000
max          99.000000
```

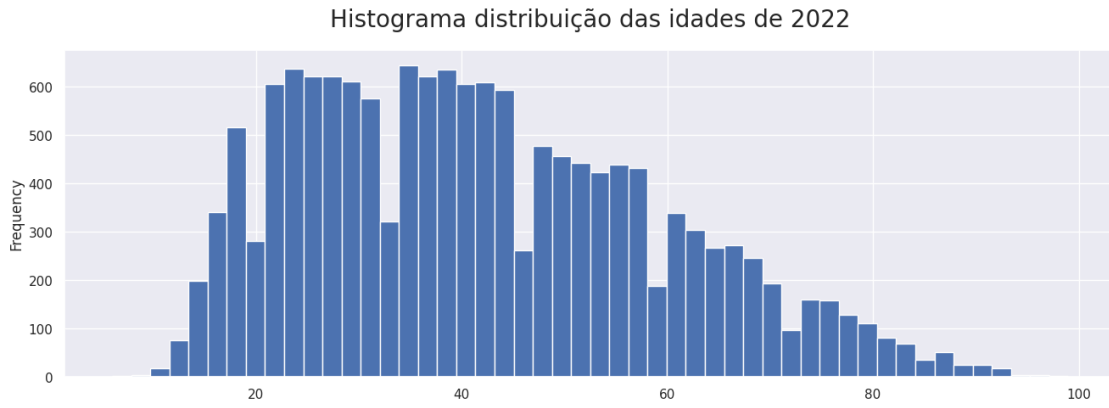
[ ]:
```
dados_2022['idade_calculada'] = (dados_2022['DTOBITO'] - dados_2022['DTNASC']) /
↪ np.timedelta64(1, 'Y')
dados_2022['idade_calculada'] = dados_2022['idade_calculada'].astype(int)
dados_2022.head()
```

[ ]:
```
       CODUFOCOR CAUSABAS      OCUP    DTOBITO HORAOBITO  NATURAL  CODMUNNATU  \
40897         27     X700       NaN 2022-01-01  10:01:00    827.0    270930.0
40898         13     X700  999991.0 2022-01-01  03:01:00    813.0    130260.0
40899         42     X709  999992.0 2022-01-01  05:01:00    841.0    410940.0
40900         31     X700       NaN 2022-01-01  14:01:00    831.0    312240.0
40901         21     X700  622020.0 2022-01-01  06:01:00    821.0    210160.0

          DTNASC  IDADE       SEXO   RACACOR         ESTCIV             ESC  \
40897 1978-04-09    443   feminino     parda       solteiro   de 1 a 3 anos
40898 1995-10-19    426  masculino     parda       solteiro  de 8 a 11 anos
40899 1962-05-28    459   feminino    branca         casado   de 1 a 3 anos
```

```
40900 1983-01-09    438   feminino    branca          viuvo   de 4 a 7 anos
40901 1993-07-22    428  masculino  indigena  uniao estavel   de 1 a 3 anos

                            ESC2010  SERIESCFAL  CODMUNRES    LOCOCOR  \
40897   fundamental I (1ª a 4ª série)         NaN     270550  domicílio
40898          médio (antigo 2º grau)         3.0     130260  domicílio
40899   fundamental I (1ª a 4ª série)         NaN     410940   hospital
40900  fundamental II (5ª a 8ª série)         NaN     310430  domicílio
40901   fundamental I (1ª a 4ª série)         1.0     210160     outros

       CODMUNOCOR ASSISTMED NECROPSIA  COMUNSVOIM  DTATESTADO CIRCOBITO  \
40897      270550       não       sim         NaN  2022-01-01  suicídio
40898      130260  ignorado  ignorado    130260.0  2022-01-01  suicídio
40899      420890       não       sim    420890.0  2022-01-01  suicídio
40900      310430       não       sim    315180.0  2022-01-01  suicídio
40901      210160       NaN       NaN         NaN  2022-02-08  suicídio

                     FONTE  TPPOS CAUSABAS_O                 ESCFALAGR1  \
40897  ocorrência policial    NaN       X709   fundamental I incompleto
40898  ocorrência policial  False       X700      ensino médio completo
40899                outra    NaN       X709   fundamental I incompleto
40900  ocorrência policial    NaN       X700  fundamental II incompleto
40901             hospital  False       X700   fundamental I incompleto

       CONTADOR                     TITULO_OCUPACAO  \
40897    510632                            ignorado
40898     17804                            ignorado
40899    309805                            ignorado
40900    775559                            ignorado
40901    446837  Trabalhador volante da agricultura

                                                CID10          UFOCOR  \
40897  Lesão autoprovocada intencionalmente por enfor…         Alagoas
40898  Lesão autoprovocada intencionalmente por enfor…        Amazonas
40899  Lesão autoprovocada intencionalmente por enfor…  Santa Catarina
40900  Lesão autoprovocada intencionalmente por enfor…    Minas Gerais
40901  Lesão autoprovocada intencionalmente por enfor…        Maranhão

        ANO  idade_calculada
40897  2022              43
40898  2022              26
40899  2022              59
40900  2022              38
40901  2022              28
```

[ ]:
```
plt.figure(figsize = (16, 5))
```

```
dados_2022.idade_calculada.plot(kind='hist', bins=50, xlabel="Anos")
plt.title(label="Histograma distribuição das idades de 2022", fontsize=20,␣
↪pad=20)
plt.show()
```



[ ]:
```
pd.DataFrame(dados_2022.idade_calculada.describe())
```

[ ]:
```
       idade_calculada
count     14830.000000
mean         41.950438
std          17.366790
min           6.000000
25%          28.000000
50%          40.000000
75%          54.000000
max          99.000000
```

Podemos observar que a estatística descritiva dos anos das vítimas não apresentou mudança percepitível. A média sempre girando nos 42 anos, com o desvio padrão em torno de 17 e mediana girando nos 40 anos.

2.9 Suicídio pela causa básica - CID-10

[ ]:
```
suicidio_por_cid = gerar_dataframe_categorico(dados.CID10, 'CID10', 'contagem')
suicidio_por_cid
```

[ ]:
```
                                                CID10  contagem
0    Lesão autoprovocada intencionalmente por enfor…     30479
1    Lesão autoprovocada intencionalmente por enfor…      4879
2    Lesão autoprovocada intencionalmente por enfor…      2205
3    Lesão autoprovocada intencionalmente por dispa…      1527
4    Lesão autoprovocada intencionalmente por enfor…      1480
..                                                 …         …
```

```
192  Lesão autoprovocada intencionalmente por meios…         1
193  Lesão autoprovocada intencionalmente por objet…         1
194  Agressão por meio de disparo de outra arma de …         1
195  Auto-intoxicação voluntária por álcool - habit…         1
196  Lesão autoprovocada intencionalmente por afoga…         1

[197 rows x 2 columns]
```

[ ]:
```
dados.CID10.value_counts(normalize=True)
```

[ ]:
```
Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e
sufocação - residência                          0.546934
Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e
sufocação - local não especificado              0.087552
Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e
sufocação - outros locais especificados         0.039568
Lesão autoprovocada intencionalmente por disparo de outra arma de fogo e de arma
de fogo não especificada - residência    0.027401
Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e
sufocação - rua e estrada                       0.026558
                                                …
Lesão autoprovocada intencionalmente por meios não especificados - área para a
prática de esportes e atletismo            0.000018
Lesão autoprovocada intencionalmente por objeto cortante ou penetrante - áreas
industriais e em construção                0.000018
Agressão por meio de disparo de outra arma de fogo ou de arma não especificada -
residência                                 0.000018
Auto-intoxicação voluntária por álcool - habitação coletiva
0.000018
Lesão autoprovocada intencionalmente por afogamento e submersão - área para a
prática de esportes e atletismo             0.000018
Name: CID10, Length: 197, dtype: float64
```

[ ]:
```
dados.CID10.value_counts(normalize=True).head(6)*100
```

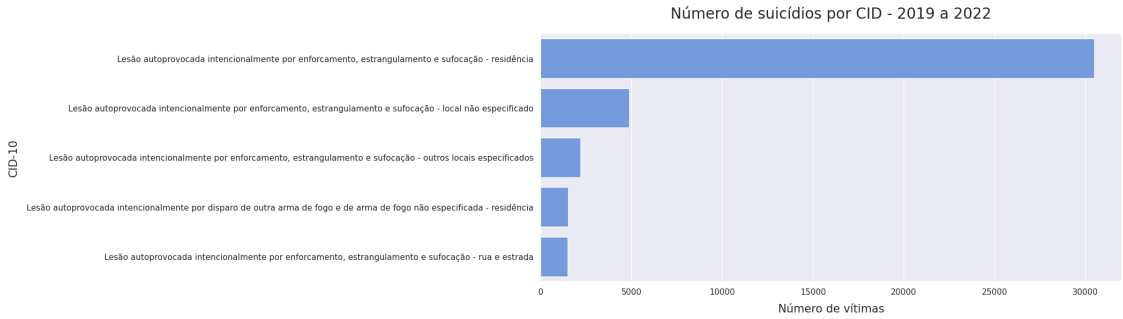
[ ]:
```
Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e
sufocação - residência                          54.693416
Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e
sufocação - local não especificado               8.755182
Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e
sufocação - outros locais especificados           3.956789
Lesão autoprovocada intencionalmente por disparo de outra arma de fogo e de arma
de fogo não especificada - residência     2.740144
Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e
sufocação - rua e estrada                         2.655804
Lesão autoprovocada intencionalmente por precipitação de um lugar elevado -
residência                                      1.613222
```

```
Name: CID10, dtype: float64
```

[ ]:
```
dados.CID10.value_counts(normalize=True)[:3].sum()
```

[ ]: 0.674053869757927

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'CID10', suicidio_por_cid.head(5),
                  'Número de suicídios por CID - 2019 a 2022', 'Número de␣
 ↪vítimas', 'CID-10')
```



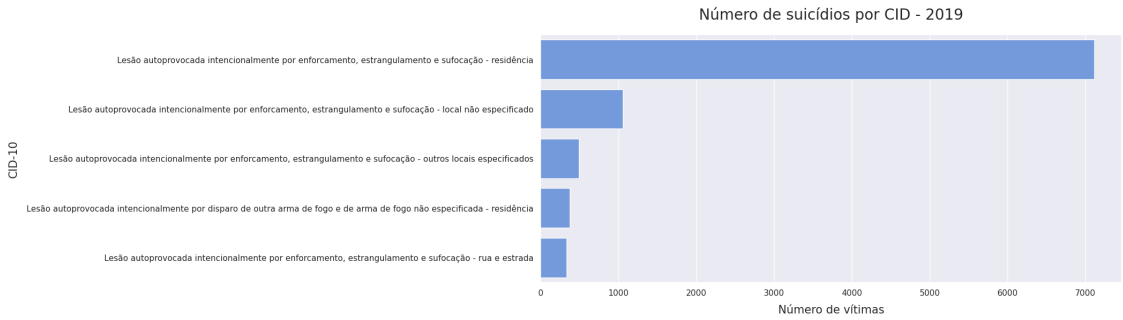
Destaque não esperado para a Lesão auto provocada intencionalmente por enforcamento, etc na residência.

Outro ponto a se destacar é que o suicidio por arma de fogo é o segundo método comparado com o enforcamento independente do local escolhido.

[ ]:
```
suicidio_por_cid_19 = gerar_dataframe_categorico(dados_2019.CID10, 'CID10',␣
 ↪'contagem')
suicidio_por_cid_19
```

[ ]:
```
                                                CID10  contagem
0    Lesão autoprovocada intencionalmente por enfor…      7113
1    Lesão autoprovocada intencionalmente por enfor…      1056
2    Lesão autoprovocada intencionalmente por enfor…       490
3    Lesão autoprovocada intencionalmente por dispa…       376
4    Lesão autoprovocada intencionalmente por enfor…       332
..                                                 …         …
144  Lesão autoprovocada intencionalmente por outro…         1
145  Auto-intoxicação por e exposição, intencional,…         1
146  Lesão autoprovocada intencionalmente por meios…         1
147  Lesão autoprovocada intencionalmente por dispo…         1
148  Auto-intoxicação voluntária por álcool - áreas…         1

[149 rows x 2 columns]
```

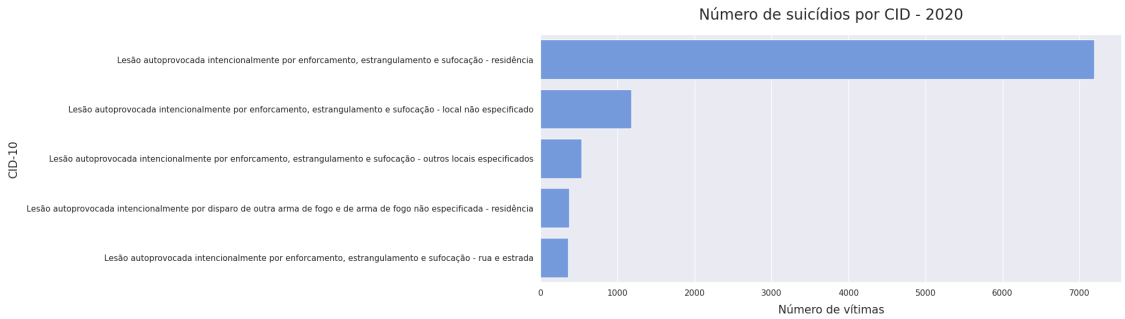
[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'CID10', suicidio_por_cid_19.head(5),
                  'Número de suicídios por CID - 2019', 'Número de vítimas',␣
 ↪'CID-10')
```



[ ]:
```
suicidio_por_cid_20 = gerar_dataframe_categorico(dados_2020.CID10, 'CID10',␣
 ↪'contagem')
suicidio_por_cid_20
```

[ ]:
```
                                                CID10  contagem
0    Lesão autoprovocada intencionalmente por enfor…     7191
1    Lesão autoprovocada intencionalmente por enfor…     1183
2    Lesão autoprovocada intencionalmente por enfor…      535
3    Lesão autoprovocada intencionalmente por dispa…      373
4    Lesão autoprovocada intencionalmente por enfor…      359
..                                                 …        …
147  Lesão autoprovocada intencionalmente por dispa…        1
148  Auto-intoxicação por e exposição, intencional,…        1
149  Lesão autoprovocada intencionalmente por meios…        1
150  Lesão autoprovocada intencionalmente pela fuma…        1
151  Auto-intoxicação intencional por outros gases …        1

[152 rows x 2 columns]
```

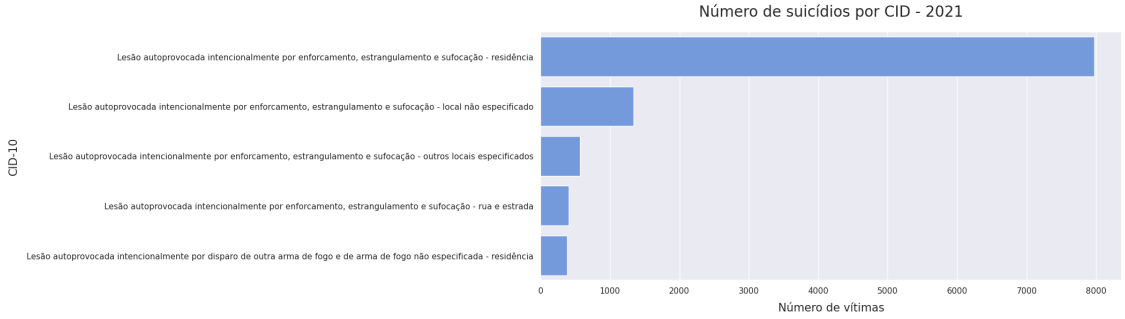
[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_barras('contagem', 'CID10', suicidio_por_cid_20.head(5),
                  'Número de suicídios por CID - 2020', 'Número de vítimas',␣
 ↪'CID-10')
```




```
[ ]: suicidio_por_cid_21 = gerar_dataframe_categorico(dados_2021.CID10, 'CID10',␣
     ↪'contagem')
     suicidio_por_cid_21
```

```
[ ]:                                                CID10  contagem
     0    Lesão autoprovocada intencionalmente por enfor…      7972
     1    Lesão autoprovocada intencionalmente por enfor…      1339
     2    Lesão autoprovocada intencionalmente por enfor…       569
     3    Lesão autoprovocada intencionalmente por enfor…       404
     4    Lesão autoprovocada intencionalmente por dispa…       385
     ..                                                  …        …
     142  Auto-intoxicação por e exposição, intencional,…         1
     143  Lesão autoprovocada intencionalmente por impac…         1
     144  Lesão autoprovocada intencionalmente por objet…         1
     145  Lesão autoprovocada intencionalmente por preci…         1
     146  Lesão autoprovocada intencionalmente por dispa…         1

     [147 rows x 2 columns]
```

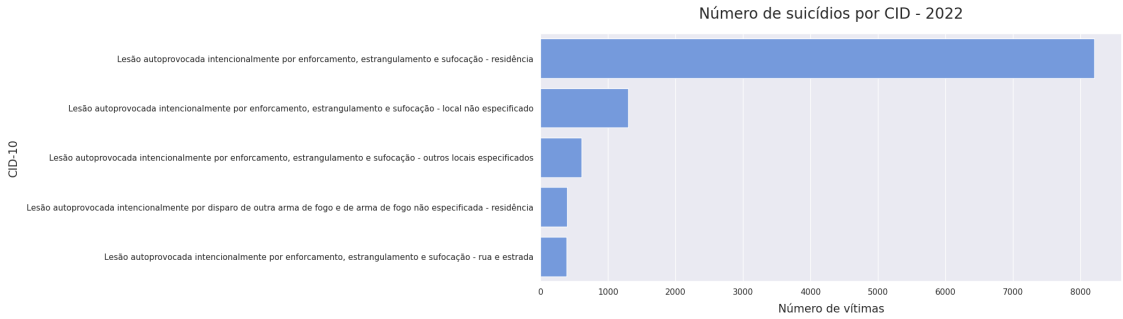
```
[ ]: #gráfico de colunas conforme função personalizada
     grafico_de_barras('contagem', 'CID10', suicidio_por_cid_21.head(5),
                       'Número de suicídios por CID - 2021', 'Número de vítimas',␣
     ↪'CID-10')
```



```
[ ]: suicidio_por_cid_22 = gerar_dataframe_categorico(dados_2022.CID10, 'CID10',␣
     ↪'contagem')
     suicidio_por_cid_22
```

```
[ ]:                                                CID10  contagem
     0    Lesão autoprovocada intencionalmente por enfor…      8203
     1    Lesão autoprovocada intencionalmente por enfor…      1301
     2    Lesão autoprovocada intencionalmente por enfor…       611
     3    Lesão autoprovocada intencionalmente por dispa…       393
     4    Lesão autoprovocada intencionalmente por enfor…       385
     ..                                                  …       …
     148  Lesão autoprovocada intencionalmente pela fuma…         1
     149  Fatos ou eventos não especificados e intenção …         1
     150  Lesão autoprovocada intencionalmente por vapor…         1
     151  Lesão autoprovocada intencionalmente por impac…         1
     152  Auto-intoxicação por e exposição, intencional,…         1

     [153 rows x 2 columns]
```

```
[ ]: #gráfico de colunas conforme função personalizada
     grafico_de_barras('contagem', 'CID10', suicidio_por_cid_22.head(5),
                       'Número de suicídios por CID - 2022', 'Número de vítimas',␣
     ↪'CID-10')
```



Ao analizar o CID-10 dos anos de 2019 a 2022 separadamente para o somatório, não encontrou-se diferença perceptível.

2.9.1 CID por sexo

```
[ ]: suicidio_cid_sexo = dados[['CID10', 'SEXO', 'DTOBITO']].groupby(['CID10',␣
     ↪'SEXO']).count().sort_values('DTOBITO', ascending=False).head(30)
     suicidio_cid_sexo
```

```
[ ]:                                                          DTOBITO
     CID10                                            SEXO
```

```
Lesão autoprovocada intencionalmente por enforc… masculino    24523
                                                  feminino      5953
Lesão autoprovocada intencionalmente por enforc… masculino     4055
Lesão autoprovocada intencionalmente por enforc… masculino     1979
Lesão autoprovocada intencionalmente por dispar… masculino     1363
Lesão autoprovocada intencionalmente por enforc… masculino     1341
Lesão autoprovocada intencionalmente por enforc… feminino       824
Lesão autoprovocada intencionalmente por dispar… masculino      669
Lesão autoprovocada intencionalmente por enforc… masculino      527
Lesão autoprovocada intencionalmente por enforc… masculino      521
Lesão autoprovocada intencionalmente por precip… masculino      492
Auto-intoxicação por e exposição, intencional, … feminino       464
Lesão autoprovocada intencionalmente por objeto… masculino      454
Lesão autoprovocada intencionalmente por dispar… masculino      432
Lesão autoprovocada intencionalmente por precip… feminino       407
Auto-intoxicação por e exposição, intencional, … masculino      401
Auto-intoxicação por e exposição, intencional, … feminino       397
Auto-intoxicação por e exposição, intencional, … masculino      346
Lesão autoprovocada intencionalmente por enforc… masculino      346
Auto-intoxicação por e exposição, intencional, … masculino      337
Auto-intoxicação por e exposição, intencional, … feminino       323
Lesão autoprovocada intencionalmente por precip… masculino      289
Auto-intoxicação por e exposição, intencional, … masculino      286
Lesão autoprovocada intencionalmente por meios … masculino      276
Lesão autoprovocada intencionalmente por precip… masculino      256
Lesão autoprovocada intencionalmente por meios … masculino      241
Auto-intoxicação por e exposição, intencional, … feminino       233
Lesão autoprovocada intencionalmente por enforc… feminino       225
Lesão autoprovocada intencionalmente por enforc… masculino      218
Lesão autoprovocada intencionalmente por dispar… masculino      203
```

[ ]:
```
suicidio_cid_sexo[:15]
```

[ ]:
```
                                                          DTOBITO
CID10                                                SEXO
Lesão autoprovocada intencionalmente por enforc… masculino    24523
                                                  feminino      5953
Lesão autoprovocada intencionalmente por enforc… masculino     4055
Lesão autoprovocada intencionalmente por enforc… masculino     1979
Lesão autoprovocada intencionalmente por dispar… masculino     1363
Lesão autoprovocada intencionalmente por enforc… masculino     1341
Lesão autoprovocada intencionalmente por enforc… feminino       824
Lesão autoprovocada intencionalmente por dispar… masculino      669
Lesão autoprovocada intencionalmente por enforc… masculino      527
Lesão autoprovocada intencionalmente por enforc… masculino      521
Lesão autoprovocada intencionalmente por precip… masculino      492
Auto-intoxicação por e exposição, intencional, … feminino       464
```

```
Lesão autoprovocada intencionalmente por objeto… masculino      454
Lesão autoprovocada intencionalmente por dispar… masculino      432
Lesão autoprovocada intencionalmente por precip… feminino       407
```

Podemos destacar neste ponto que há uma diferença do método que o sexo escolhe para cometer o suicídio.

[ ]:
```
suicidio_cid_sexo.loc['Lesão autoprovocada intencionalmente por disparo de arma␣
↪de fogo de mão - residência']
```

[ ]:
```
           DTOBITO
SEXO
masculino      669
```

[ ]:
```
suicidio_cid_sexo.loc['Lesão autoprovocada intencionalmente por precipitação de␣
↪um lugar elevado - residência']
```

[ ]:
```
           DTOBITO
SEXO
masculino      492
feminino       407
```

[ ]:
```
suicidio_cid_sexo.loc['Auto-intoxicação por e exposição, intencional, a outras␣
↪drogas, medicamentos e substâncias biológicas e às não especificadas -␣
↪residência']
```

[ ]:
```
           DTOBITO
SEXO
feminino       464
masculino      346
```

Podemos perceber uma preferência de métodos quanto ao sexo.

2.10 Suicídio por estado civil

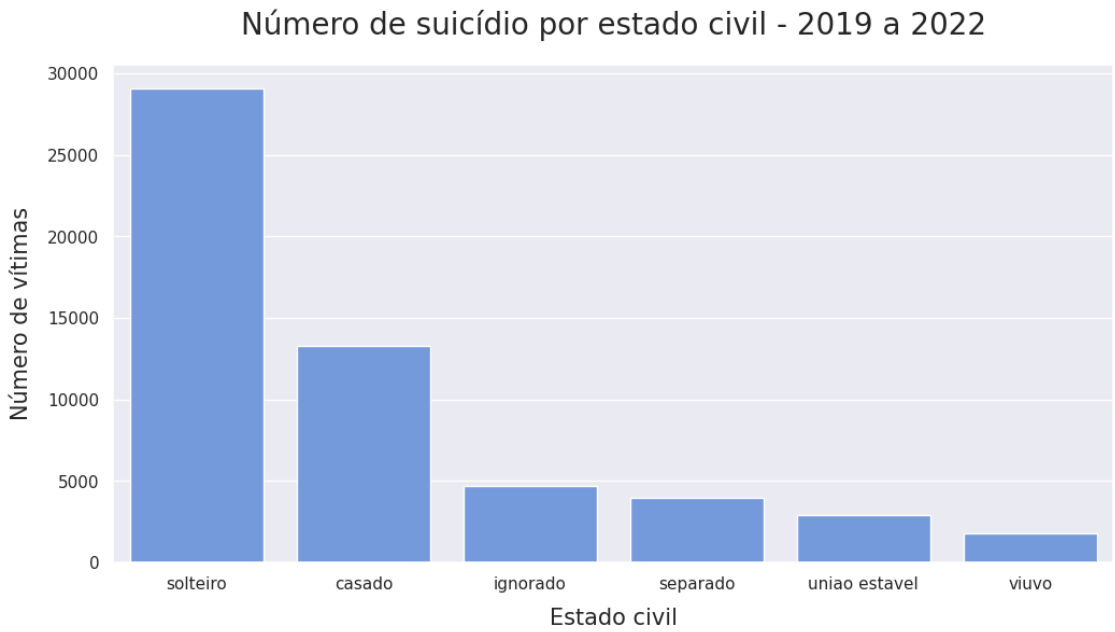
[ ]:
```
suicidio_por_estado_civil = gerar_dataframe_categorico(dados.ESTCIV, 'ESTCIV',␣
↪'contagem')
suicidio_por_estado_civil
```

[ ]:
```
          ESTCIV  contagem
0       solteiro     29082
1         casado     13274
2       ignorado      4681
3       separado      3966
4  uniao estavel      2914
5          viuvo      1810
```

[ ]:
```
dados.ESTCIV.value_counts(normalize=True)
```

[ ]:
```
solteiro         0.521866
casado           0.238197
ignorado         0.083999
separado         0.071168
uniao estavel    0.052291
viuvo            0.032480
Name: ESTCIV, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('ESTCIV', 'contagem', suicidio_por_estado_civil,
                   'Número de suicídio por estado civil - 2019 a 2022', 'Estado␣
 ↪civil', 'Número de vítimas')
```



[ ]:
```
suicidio_por_estado_civil_19 = gerar_dataframe_categorico(dados_2019.ESTCIV,␣
 ↪'ESTCIV', 'contagem')
suicidio_por_estado_civil_19
```

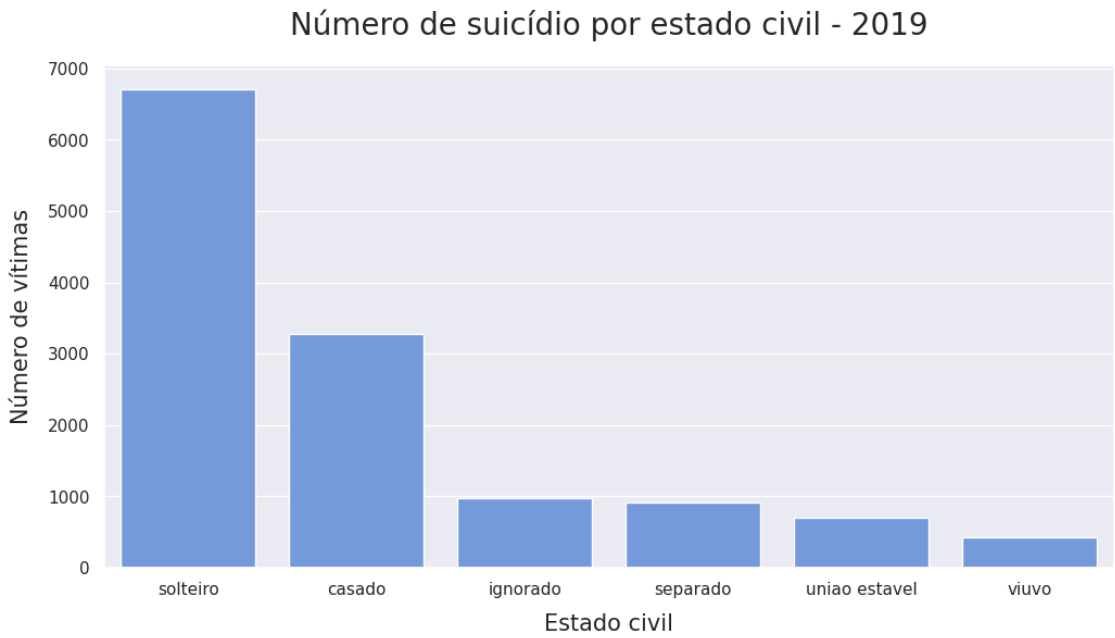
[ ]:

	ESTCIV	contagem
0	solteiro	6707
1	casado	3282
2	ignorado	981
3	separado	916
4	uniao estavel	702
5	viuvo	421

[ ]:
```
dados_2019.ESTCIV.value_counts(normalize=True)
```

[ ]:
```
solteiro         0.515566
casado           0.252287
ignorado         0.075409
separado         0.070413
uniao estavel    0.053963
viuvo            0.032362
Name: ESTCIV, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('ESTCIV', 'contagem', suicidio_por_estado_civil_19,
                   'Número de suicídio por estado civil - 2019', 'Estado␣
 ↪civil', 'Número de vítimas')
```



[ ]:
```
suicidio_por_estado_civil_20 = gerar_dataframe_categorico(dados_2020.ESTCIV,␣
 ↪'ESTCIV', 'contagem')
suicidio_por_estado_civil_20
```

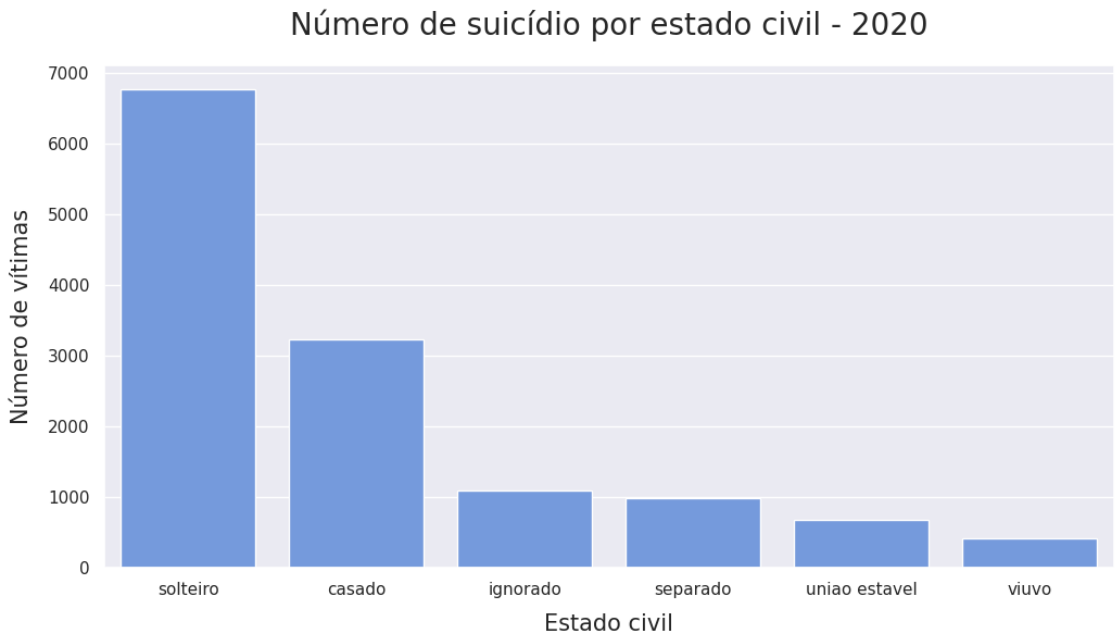
[ ]:

	ESTCIV	contagem
0	solteiro	6765
1	casado	3232
2	ignorado	1095
3	separado	978
4	uniao estavel	676
5	viuvo	420

[ ]:
```
dados_2020.ESTCIV.value_counts(normalize=True)
```

[ ]:
```
solteiro         0.513823
casado           0.245481
ignorado         0.083169
separado         0.074282
uniao estavel    0.051344
viuvo            0.031900
Name: ESTCIV, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('ESTCIV', 'contagem', suicidio_por_estado_civil_20,
                   'Número de suicídio por estado civil - 2020', 'Estado␣
 ↪civil', 'Número de vítimas')
```



[ ]:
```
suicidio_por_estado_civil_21 = gerar_dataframe_categorico(dados_2021.ESTCIV,␣
 ↪'ESTCIV', 'contagem')
suicidio_por_estado_civil_21
```

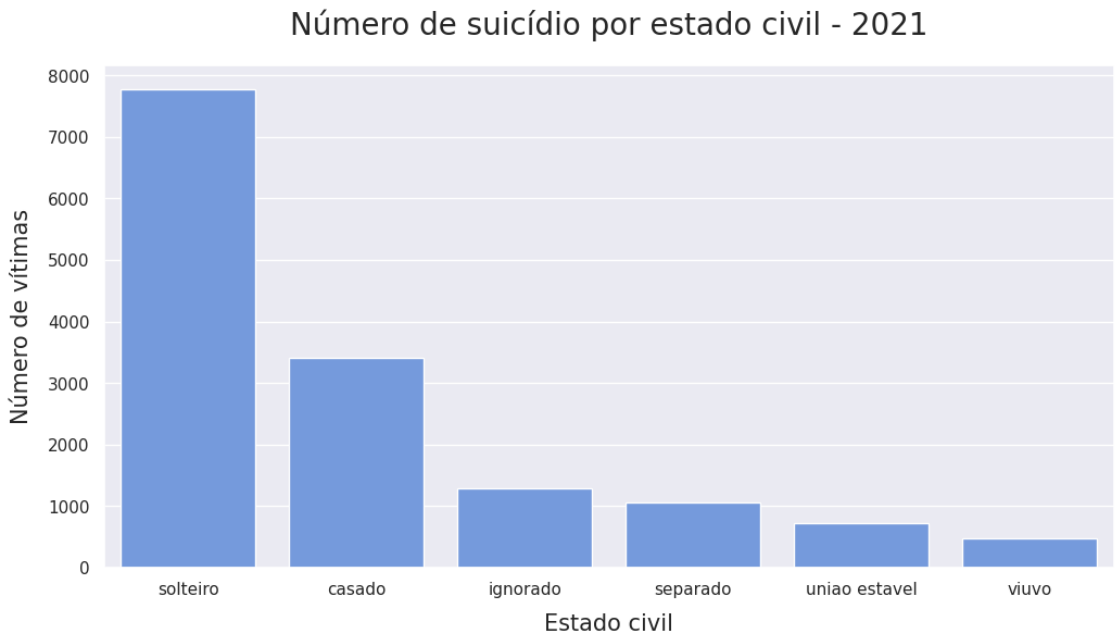
[ ]:

	ESTCIV	contagem
0	solteiro	7775
1	casado	3407
2	ignorado	1284
3	separado	1056
4	uniao estavel	719
5	viuvo	481

[ ]:
```
dados_2021.ESTCIV.value_counts(normalize=True)
```

[ ]:
```
solteiro         0.528121
casado           0.231422
ignorado         0.087216
separado         0.071729
uniao estavel    0.048838
viuvo            0.032672
Name: ESTCIV, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('ESTCIV', 'contagem', suicidio_por_estado_civil_21,
                   'Número de suicídio por estado civil - 2021', 'Estado␣
 ↪civil', 'Número de vítimas')
```



[ ]:
```
suicidio_por_estado_civil_22 = gerar_dataframe_categorico(dados_2022.ESTCIV,␣
 ↪'ESTCIV', 'contagem')
suicidio_por_estado_civil_22
```

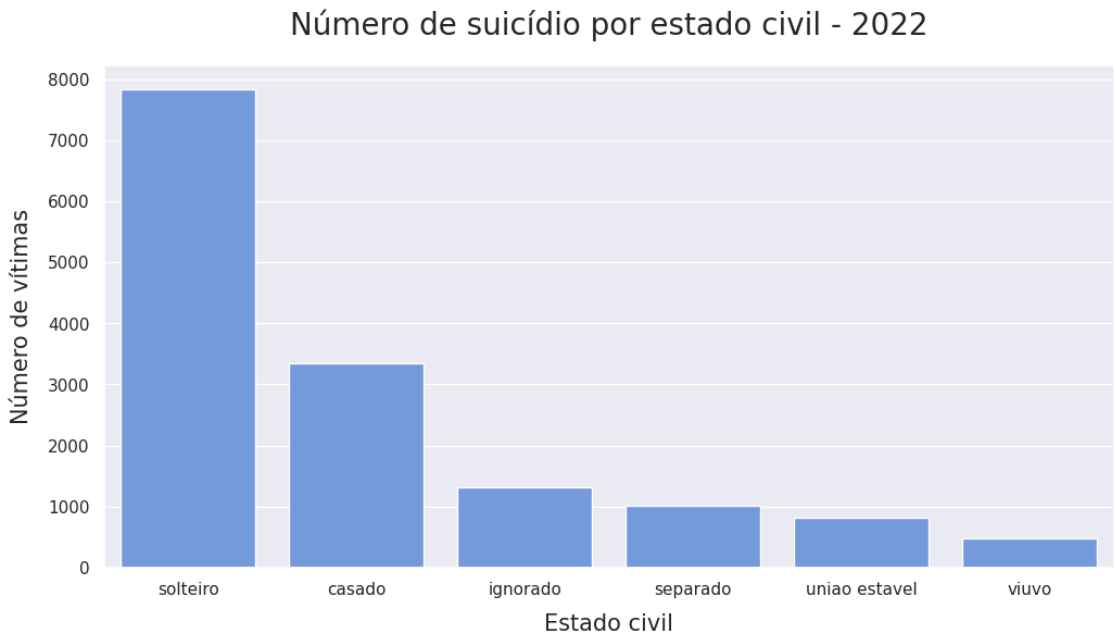
[ ]:

	ESTCIV	contagem
0	solteiro	7835
1	casado	3353
2	ignorado	1321
3	separado	1016
4	uniao estavel	817
5	viuvo	488

[ ]:
```
dados_2022.ESTCIV.value_counts(normalize=True)
```

[ ]:
```
solteiro         0.528321
casado           0.226096
ignorado         0.089076
separado         0.068510
uniao estavel    0.055091
viuvo            0.032906
Name: ESTCIV, dtype: float64
```

[ ]:
```
#gráfico de colunas conforme função personalizada
grafico_de_colunas('ESTCIV', 'contagem', suicidio_por_estado_civil_22,
                   'Número de suicídio por estado civil - 2022', 'Estado␣
 ↪civil', 'Número de vítimas')
```



2.11 Suicídio por estado civil por sexo

[ ]:
```
dados[['ESTCIV', 'SEXO', 'DTOBITO']].groupby(['SEXO', 'ESTCIV']).count().
 ↪sort_values(by=['DTOBITO', 'SEXO'], ascending=False).unstack(fill_value=0)
```

[ ]:

	DTOBITO					
ESTCIV	casado	ignorado	separado	solteiro	uniao estavel	viuvo
SEXO						
feminino	2672	884	1070	6200	592	666
ignorado	0	9	0	0	1	0
masculino	10602	3788	2896	22882	2321	1144

[ ]:
```
dados[['ESTCIV', 'SEXO', 'DTOBITO']].groupby(['SEXO', 'ESTCIV']).count().
↪sort_values(by=['SEXO', 'DTOBITO'], ascending=False)
```

[ ]:
```
                          DTOBITO
SEXO      ESTCIV
masculino solteiro         22882
          casado           10602
          ignorado          3788
          separado          2896
          uniao estavel     2321
          viuvo             1144
ignorado  ignorado             9
          uniao estavel        1
feminino  solteiro          6200
          casado            2672
          separado          1070
          ignorado           884
          viuvo              666
          uniao estavel      592
```

2.12 Suicídio por estado civil e raça

[ ]:
```
dados[['ESTCIV', 'RACACOR', 'DTOBITO']].groupby(['RACACOR', 'ESTCIV']).count().
↪sort_values(by=['RACACOR', 'DTOBITO'], ascending=False).unstack(fill_value=0)
```

[ ]:
```
         DTOBITO
ESTCIV    casado ignorado separado solteiro uniao estavel viuvo
RACACOR
amarela       49       14       13       75             7     9
branca      7626     2187     2482    12089          1167  1091
ignorado      87      303       37      213            28    15
indigena      34       58        3      374            53     4
parda       4957     1890     1300    14484          1497   615
preta        521      229      131     1847           162    76
```

Profile report – DOEXT19

From Cloud Storage to BigQuery – Job ID: 21535139

All Data

31 columns 142,800 rows 4 data types

- 93% valid values
- 1% mismatching values
- 6% missing values

Output destinations

Name	Location	Status
causas_externas_19	bigquery://mvp-puc-399114/obitos_datasus/causas_externas_19	-

Data sources

Name	Location	Schema validation
DOEXT19.csv	gs://obitos-datasus/DOEXT19.csv	No schema changes found
ibge_codigo_UF.csv	gs://dados-complementares/ibge_codigo_UF.csv	No schema changes found
tabela_cid10.csv	gs://dados-complementares/tabela_cid10.csv	
tabela_ocupacoes.csv	gs://dados-complementares/tabela_ocupacoes.csv	No schema changes found

Execution summary

Job Status
Completed

Job Type
Manual

Environment
Dataflow

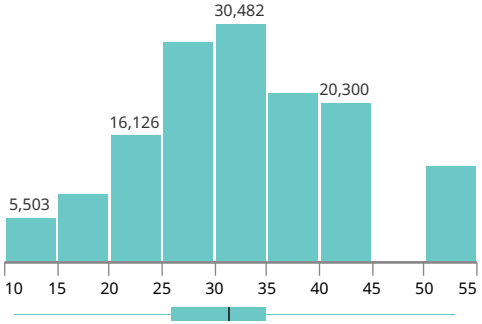
Start time
Tue, Sep 26, 2023 1:56 PM -03:00

End time
Tue, Sep 26, 2023 2:18 PM -03:00

Duration
22 min

CODUFOCOR

Type	Integer
Valid	142,800
Mismatched	0
Empty	0

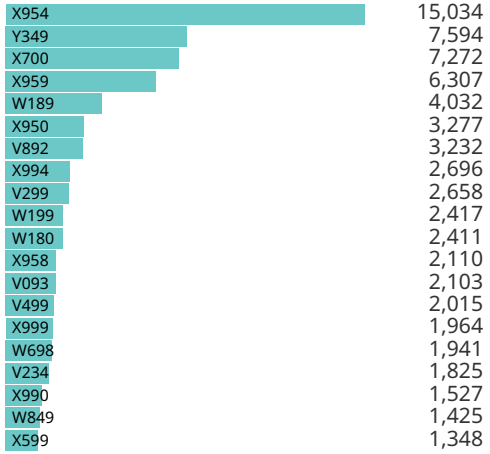


Minimum	11
Lower quartile	26
Median	31
Upper quartile	35
Maximum	53

CAUSABAS

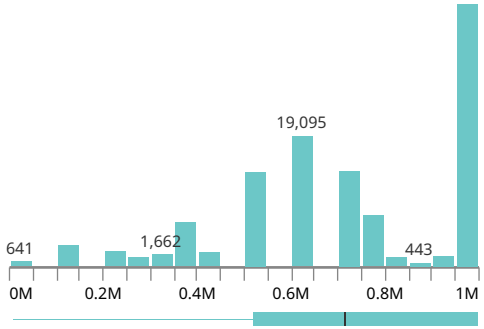
Type	String
Valid	142,800
Mismatched	0
Empty	0

Top 20 values



OCUP

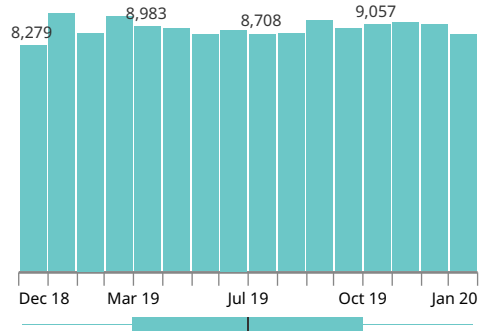
Type	Integer
Valid	113,257
Mismatched	0
Empty	29,543



Minimum	10,205
Lower quartile	520,844
Median	715,556
Upper quartile	999,991
Maximum	999,995

DTOBITO

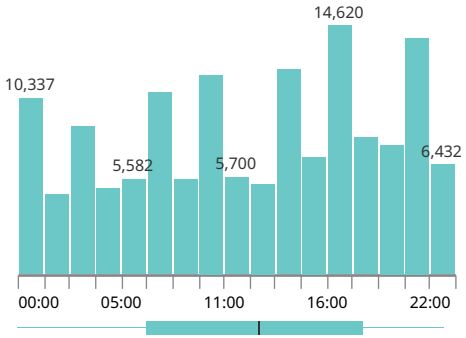
Type	Date / Time (mm*dd*yyyy)
Valid	142,800
Mismatched	0
Empty	0



Minimum	Jan 1 2019
Lower quartile	Mar 31 2019
Median	Jul 1 2019
Upper quartile	Oct 3 2019
Maximum	Dec 31 2019

HORAOBITO

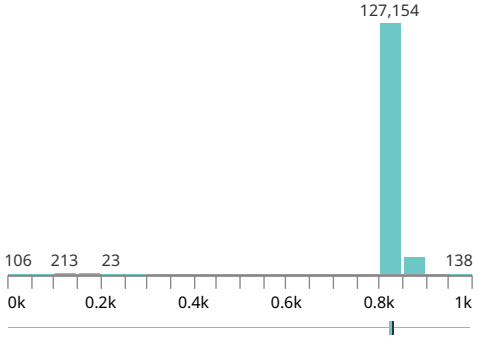
Type	Date / Time (HH:MM)
Valid	142,800
Mismatched	0
Empty	0



Minimum	00:01:00
Lower quartile	06:59:29
Median	13:01:51
Upper quartile	18:40:50
Maximum	23:01:00

NATURAL

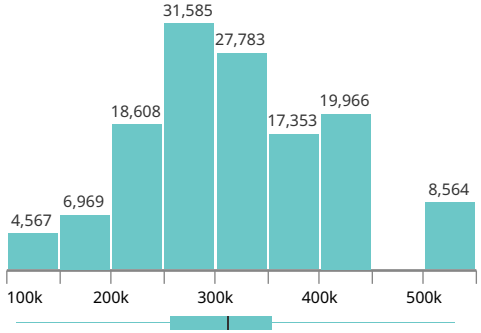
Type	Integer
Valid	136,714
Mismatched	0
Empty	6,086



Minimum	4
Lower quartile	825
Median	831
Upper quartile	835
Maximum	999

CODMUNNATU

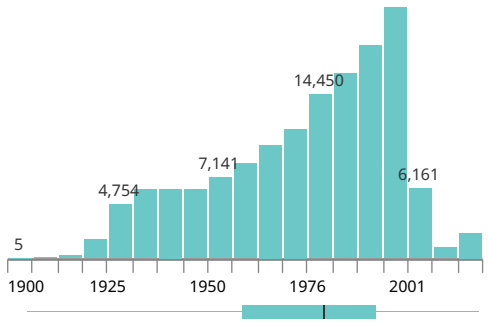
Type	Integer
Valid	135,395
Mismatched	0
Empty	7,405



Minimum	110,000
Lower quartile	257,411
Median	312,331
Upper quartile	354,967
Maximum	530,010

DTNASC

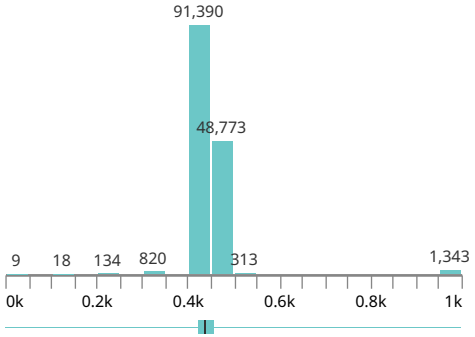
Type	Date / Time (mm*dd*yyyy)
Valid	142,800
Mismatched	0
Empty	0



Minimum	Jun 24 1905
Lower quartile	Mar 25 1960
Median	Oct 13 1980
Upper quartile	Jan 21 1994
Maximum	Dec 24 2019

IDADE

Type	Integer
Valid	142,800
Mismatched	0
Empty	0



Minimum	1
Lower quartile	425
Median	439
Upper quartile	460
Maximum	999

SEXO

Type	String
Valid	142,800
Mismatched	0
Empty	0

Top 3 values

masculino	114,469
feminino	28,136
ignorado	195

RACACOR

Type	String
Valid	142,800
Mismatched	0
Empty	0

Top 6 values

parda	76,819
branca	52,691
preta	9,735
ignorado	2,440
indigena	654
amarela	461

ESTCIV

Type	String
Valid	142,800
Mismatched	0
Empty	0

Top 6 values

solteiro	76,092
casado	28,201
ignorado	13,470
viuvo	10,848
uniao estavel	7,201
separado	6,988

ESC

Type	String
Valid	142,800
Mismatched	0
Empty	0

Top 6 values

de 4 a 7 anos	40,231
de 8 a 11 anos	33,706
ignorado	29,917
de 1 a 3 anos	22,046
nenhuma	9,725
12 anos e mais	7,175

ESC2010

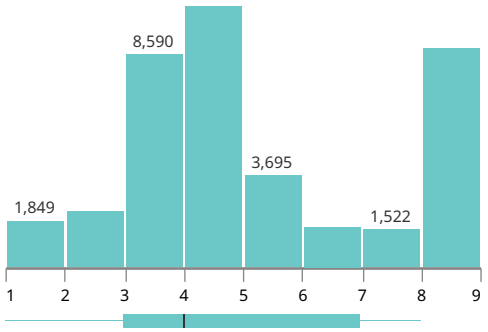
Type	String
Valid	142,800
Mismatched	0
Empty	0

Top 7 values

fundamental II (5ª a 8ª série...	38,263
fundamental I (1ª a 4ª série)	32,297
ignorado	30,715
médio (antigo 2º grau)	24,773
sem escolaridade	9,608
superior completo	4,874
superior incompleto	2,270

SERIESCFAL

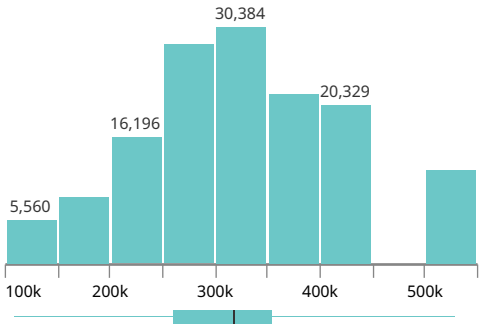
Type	Integer
Valid	38,914
Mismatched	0
Empty	103,886



Minimum	1
Lower quartile	3
Median	4
Upper quartile	7
Maximum	8

CODMUNRES

Type	Integer
Valid	142,800
Mismatched	0
Empty	0



Minimum	110,000
Lower quartile	260,695
Median	318,570
Upper quartile	355,040
Maximum	530,010

LOCOCOR

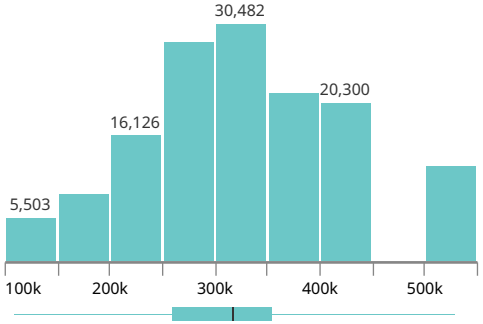
Type	String
Valid	142,800
Mismatched	0
Empty	0

Top 6 values

hospital	54,721
via pública	40,479
domicílio	21,704
outros	21,289
outros estab. de saúde	4,229
ignorado	378

CODMUNOCOR

Type	Integer
Valid	142,800
Mismatched	0
Empty	0



Minimum	110,001
Lower quartile	260,669
Median	318,124
Upper quartile	355,063
Maximum	530,010

ASSISTMED

Type	String
Valid	107,847
Mismatched	0
Empty	34,953

Top 3 values

não	54,921
sim	44,241
ignorado	8,685

NECROPSIA

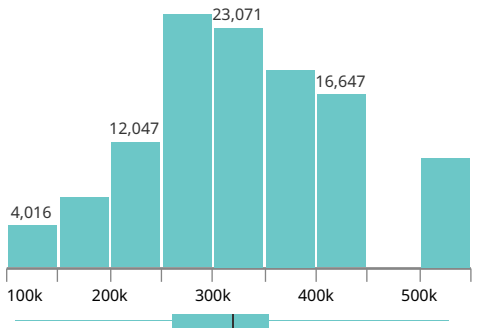
Type	String
Valid	119,252
Mismatched	0
Empty	23,548

Top 3 values

sim	98,799
não	18,597
ignorado	1,856

COMUNSVOIM

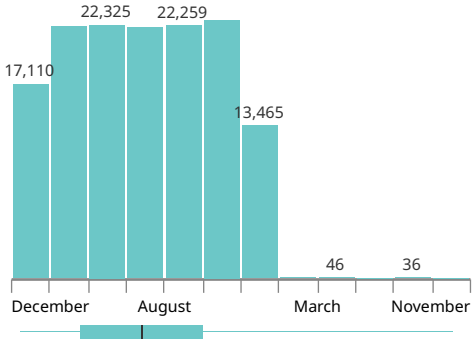
Type	Integer
Valid	116,469
Mismatched	0
Empty	26,331



Minimum	110,002
Lower quartile	261,156
Median	319,734
Upper quartile	355,173
Maximum	530,010

DTATESTADO

Type	Date / Time (mm*dd*yyyy)
Valid	142,432
Mismatched	0
Empty	368



Minimum	Jan 1 2019
Lower quartile	Apr 1 2019
Median	Jul 3 2019
Upper quartile	Oct 4 2019
Maximum	Oct 16 2020

CIRCOBITO

Type	String
Valid	142,457
Mismatched	0
Empty	343

Top 5 values

acidente	60,199
homicídio	42,033
ignorado	23,243
suicídio	13,009
outros	3,973

FONTE

Type	String
Valid	142,800
Mismatched	0
Empty	0

Top 5 values

ocorrência policial	73,625
ignorado	40,050
hospital	11,224
outra	10,882
família	7,019

TPPOS

Type	Boolean
Valid	58,999
Mismatched	37,755
Empty	46,046
Top value	
N	58,999

CAUSABAS_O

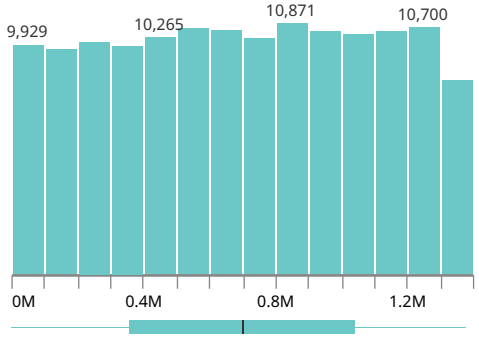
Type	String
Valid	142,744
Mismatched	0
Empty	56
Top 20 values	
X954	13,840
Y349	11,186
X700	6,756
X959	6,325
V892	3,704
X950	3,005
W189	2,848
V299	2,711
X994	2,477
W199	2,258
V499	2,131
V093	1,999
Y299	1,980
X999	1,978
R99	1,965
X599	1,898
X958	1,893
W698	1,798
W180	1,791
Y249	1,576

ESCFALAGR1

Type	String
Valid	142,800
Mismatched	0
Empty	0
Top 9 values	
ignorado	30,715
fundamental II incompleto	29,422
fundamental I incompleto	21,766
ensino médio incompleto	15,128
sem escolaridade	14,229
fundamental I completo	10,531
fundamental II completo	8,841
ensino médio completo	7,294
superior completo	4,874

CONTADOR

Type	Integer
Valid	142,800
Mismatched	0
Empty	0



Minimum	11
Lower quartile	359,835
Median	703,309
Upper quartile	1,042,559
Maximum	1,378,900

TITULO_OCUPACAO

Type	String
Valid	142,800
Mismatched	0
Empty	0
Top 20 values	
ignorado	68,168
Trabalhador agropecuário em g...	6,843
Trabalhador volante da agricu...	6,555
Representante comercial autôn...	5,658
Pedreiro	5,534
Empregado doméstico nos ser...	3,242
Servente de obras	3,000
Comerciante varejista	1,832
Produtor agrícola polivalente	1,729
Caseiro (agricultura)	1,493
Motorista de carro de passeio	1,463
Vendedor de comércio varejist...	1,349
Motorista de caminhão (rotas ...	1,161
Pintor de obras	1,120
Mecânico de manutenção de aut...	976
Vigilante	869
Motofretista	664
Vendedor ambulante	628
Eletricista de instalações	616
Motorista de táxi	529

CID10

Type	String
Valid	142,800
Mismatched	0
Empty	0
Top 20 values	
Agressão por meio de disparo ...	15,034
Fatos ou eventos não especifi...	7,594
Lesão autoprovocada intencion...	7,272
Agressão por meio de disparo ...	6,307
Outras quedas no mesmo nível ...	4,032
Agressão por meio de disparo ...	3,277
Pessoa traumatizada em um aci...	3,232
Agressão por meio de objeto c...	2,696
Motociclista [qualquer] traum...	2,658
Queda sem especificação - loc...	2,417
Outras quedas no mesmo nível ...	2,411
Agressão por meio de disparo ...	2,110
Pedestre traumatizado em um a...	2,103
Ocupante [qualquer] de um aut...	2,015
Agressão por meio de objeto c...	1,964
Afogamento e submersão em águ...	1,941
Motociclista traumatizado em ...	1,825
Agressão por meio de objeto c...	1,527
Riscos não especificados à re...	1,425
Exposição a fator não especif...	1,348

UFOCOR

Type	String
Valid	142,800
Mismatched	0
Empty	0
Top 20 values	
São Paulo	21,552
Rio de Janeiro	13,951
Minas Gerais	13,240
Bahia	12,856
Paraná	8,373
Pernambuco	8,203
Rio Grande do Sul	7,561
Ceará	6,779
Pará	6,524
Goiás	5,796
Maranhão	4,442
Santa Catarina	4,366
Espírito Santo	3,291
Amazonas	3,015
Mato Grosso	2,880
Paraíba	2,654
Rio Grande do Norte	2,579
Alagoas	2,412
Piauí	2,326
Sergipe	1,973

Profile report – DOEXT20

From Cloud Storage to BigQuery – Job ID: 21535123

All Data

31 columns 146,038 rows 4 data types

- 93% valid values
- 1% mismatching values
- 6% missing values

Output destinations

Name	Location	Status
causas_externas_20	bigquery://mvp-puc-399114/obitos_datasus/causas_externas_20	-

Data sources

Name	Location	Schema validation
ibge_codigo_UF.csv	gs://dados-complementares/ibge_codigo_UF.csv	No schema changes found
tabela_ocupacoes.csv	gs://dados-complementares/tabela_ocupacoes.csv	No schema changes found
tabela_cid10.csv	gs://dados-complementares/tabela_cid10.csv	
DOEXT20.csv	gs://obitos-datasus/DOEXT20.csv	No schema changes found

Execution summary

Job Status
Completed

Job Type
Manual

Environment
Dataflow

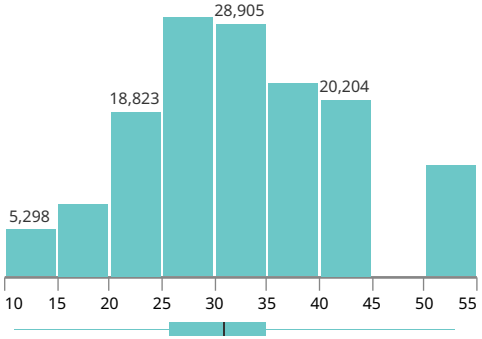
Start time
Tue, Sep 26, 2023 1:55 PM -03:00

End time
Tue, Sep 26, 2023 2:16 PM -03:00

Duration
22 min

CODUFOCOR

Type	Integer
Valid	146,038
Mismatched	0
Empty	0

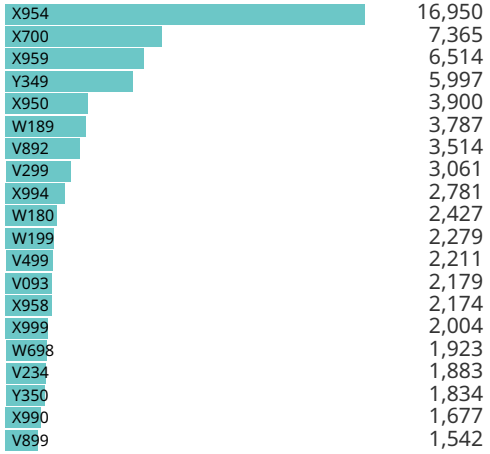


Minimum	11
Lower quartile	26
Median	31
Upper quartile	35
Maximum	53

CAUSABAS

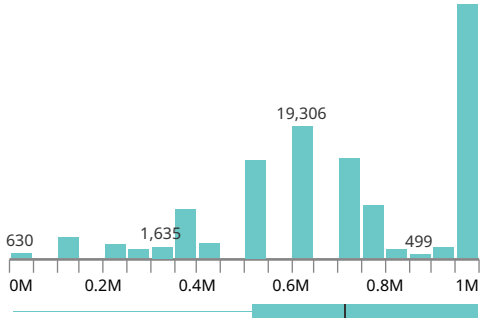
Type	String
Valid	146,038
Mismatched	0
Empty	0

Top 20 values



OCUP

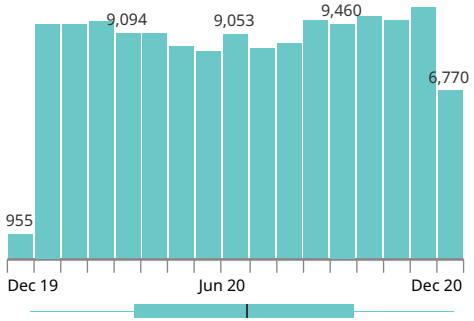
Type	Integer
Valid	114,425
Mismatched	0
Empty	31,613



Minimum	10,205
Lower quartile	519,444
Median	715,286
Upper quartile	999,991
Maximum	999,995

DTOBITO

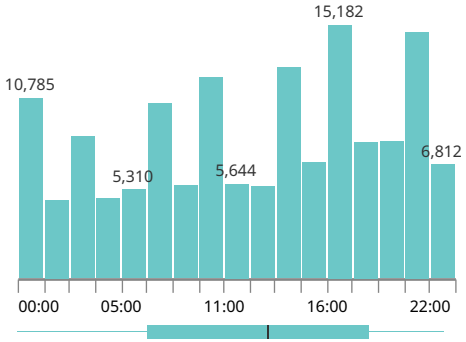
Type	Date / Time (mm*dd*yyyy)
Valid	146,038
Mismatched	0
Empty	0



Minimum	Jan 1 2020
Lower quartile	Mar 30 2020
Median	Jul 5 2020
Upper quartile	Oct 5 2020
Maximum	Dec 31 2020

HORAOBITO

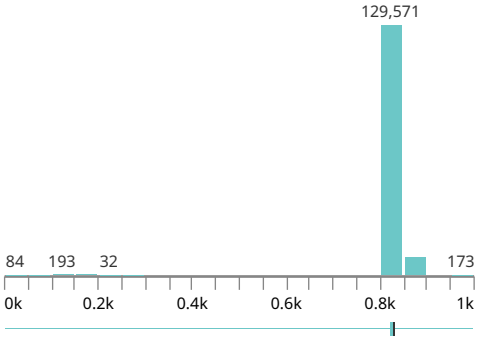
Type	Date / Time (HH:MM)
Valid	146,038
Mismatched	0
Empty	0



Minimum	00:01:00
Lower quartile	07:00:57
Median	13:28:52
Upper quartile	18:59:34
Maximum	23:01:00

NATURAL

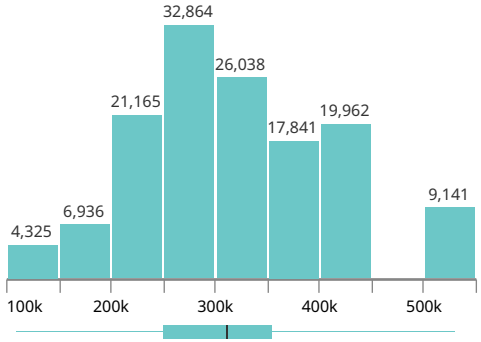
Type	Integer
Valid	139,715
Mismatched	0
Empty	6,323



Minimum	4
Lower quartile	825
Median	831
Upper quartile	835
Maximum	999

CODMUNNATU

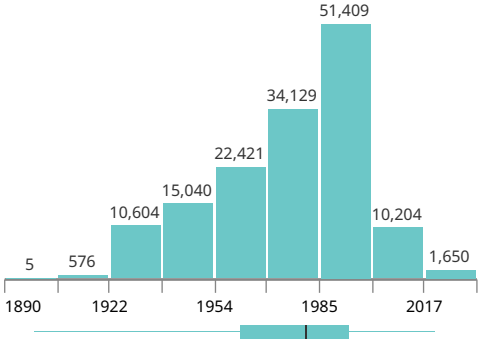
Type	Integer
Valid	138,272
Mismatched	0
Empty	7,766



Minimum	110,000
Lower quartile	251,260
Median	311,322
Upper quartile	354,953
Maximum	530,010

DTNASC

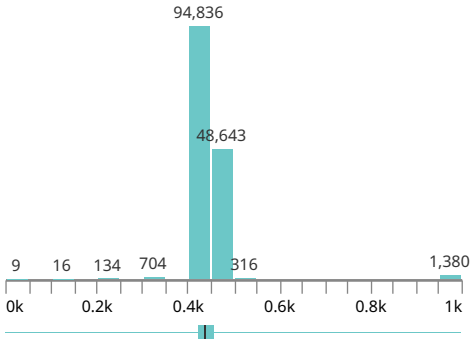
Type	Date / Time (mm*dd*yyyy)
Valid	146,038
Mismatched	0
Empty	0



Minimum	Mar 6 1900
Lower quartile	Jun 22 1962
Median	Jan 23 1982
Upper quartile	Feb 2 1995
Maximum	Dec 26 2020

IDADE

Type	Integer
Valid	146,038
Mismatched	0
Empty	0



Minimum	1
Lower quartile	425
Median	438
Upper quartile	458
Maximum	999

SEXO

Type	String
Valid	146,038
Mismatched	0
Empty	0

Top 3 values

masculino	118,362
feminino	27,460
ignorado	216

RACACOR

Type	String
Valid	146,038
Mismatched	0
Empty	0

Top 6 values

parda	80,224
branca	51,799
preta	10,052
ignorado	2,753
indigena	625
amarela	585

ESTCIV

Type	String
Valid	146,038
Mismatched	0
Empty	0

Top 6 values

solteiro	79,348
casado	27,713
ignorado	14,684
viuvo	10,063
uniao estavel	7,153
separado	7,077

ESC

Type	String
Valid	146,038
Mismatched	0
Empty	0

Top 6 values

de 4 a 7 anos	41,095
de 8 a 11 anos	37,023
ignorado	30,588
de 1 a 3 anos	21,312
nenhuma	9,003
12 anos e mais	7,017

ESC2010

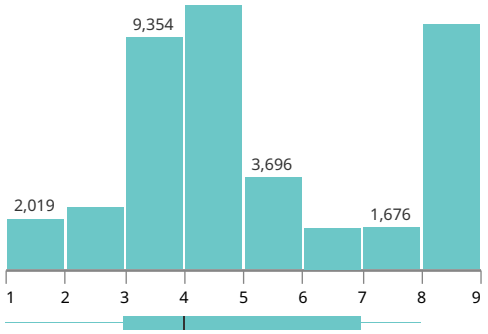
Type	String
Valid	146,038
Mismatched	0
Empty	0

Top 7 values

fundamental II (5ª a 8ª série...	40,257
fundamental I (1ª a 4ª série)	31,910
ignorado	30,811
médio (antigo 2º grau)	27,082
sem escolaridade	8,978
superior completo	4,808
superior incompleto	2,192

SERIESCFAL

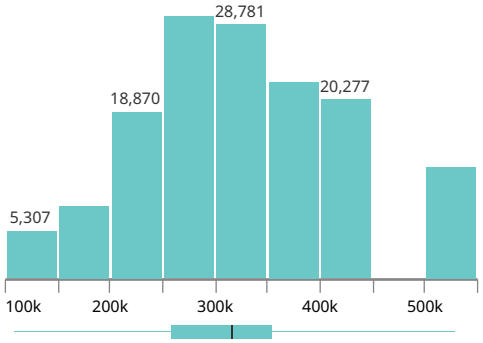
Type	Integer
Valid	41,462
Mismatched	0
Empty	104,576



Minimum	1
Lower quartile	3
Median	4
Upper quartile	7
Maximum	8

CODMUNRES

Type	Integer
Valid	146,038
Mismatched	0
Empty	0



Minimum	110,000
Lower quartile	259,035
Median	316,241
Upper quartile	355,030
Maximum	530,010

LOCOCOR

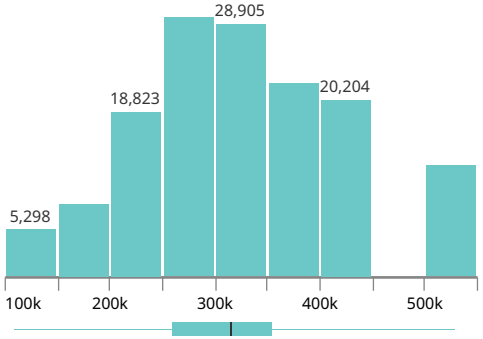
Type	String
Valid	146,038
Mismatched	0
Empty	0

Top 6 values

hospital	53,859
via pública	42,496
domicílio	23,313
outros	21,672
outros estab. de saúde	4,337
ignorado	361

CODMUNOCOR

Type	Integer
Valid	146,038
Mismatched	0
Empty	0



Minimum	110,001
Lower quartile	260,344
Median	315,988
Upper quartile	355,035
Maximum	530,010

ASSISTMED

Type	String
Valid	111,498
Mismatched	0
Empty	34,540

Top 3 values

não	58,348
sim	44,208
ignorado	8,942

NECROPSIA

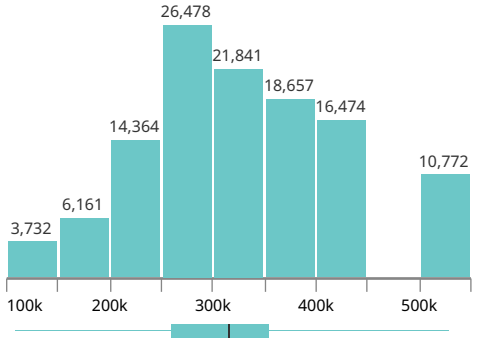
Type	String
Valid	121,576
Mismatched	0
Empty	24,462

Top 3 values

sim	95,977
não	23,320
ignorado	2,279

COMUNSVOIM

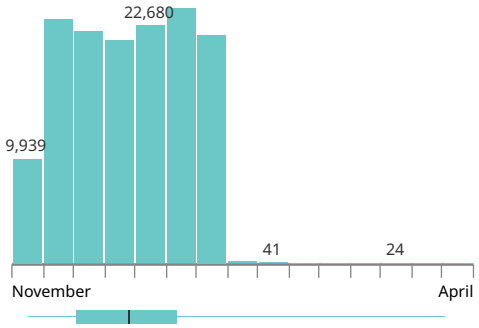
Type	Integer
Valid	118,479
Mismatched	0
Empty	27,559



Minimum	110,002
Lower quartile	261,012
Median	316,458
Upper quartile	355,095
Maximum	530,010

DTATESTADO

Type	Date / Time (mm*dd*yyyy)
Valid	145,761
Mismatched	0
Empty	277



Minimum	Jan 1 2020
Lower quartile	Apr 1 2020
Median	Jul 7 2020
Upper quartile	Oct 7 2020
Maximum	Feb 22 2022

CIRCOBITO

Type	String
Valid	145,621
Mismatched	0
Empty	417

Top 5 values

acidente	61,738
homicídio	45,681
ignorado	20,363
suicídio	13,166
outros	4,673

FONTE

Type	String
Valid	146,038
Mismatched	0
Empty	0

Top 5 values

ocorrência policial	76,344
ignorado	39,932
outra	11,346
hospital	10,951
família	7,465

TPPOS

Type	Boolean
Valid	57,295
Mismatched	39,975
Empty	48,768
Top value	
N	57,295

CAUSABAS_O

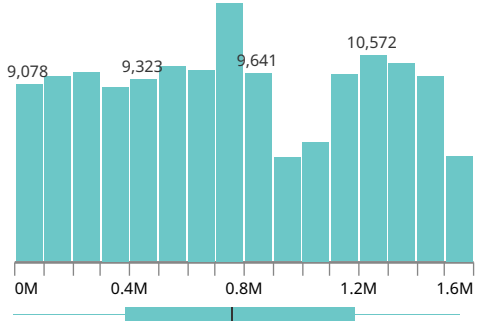
Type	String
Valid	145,704
Mismatched	0
Empty	334
Top 20 values	
X954	15,401
Y349	11,046
X700	6,730
X959	6,392
V892	4,172
X950	3,547
V299	3,050
W189	2,753
X994	2,500
R99	2,315
Y299	2,259
V499	2,234
X999	2,052
X958	2,049
W199	1,984
X599	1,919
W180	1,832
W698	1,799
V093	1,700
X990	1,531

ESCFALAGR1

Type	String
Valid	146,038
Mismatched	0
Empty	0
Top 9 values	
ignorado	30,811
fundamental II incompleto	30,358
fundamental I incompleto	21,250
ensino médio incompleto	16,326
sem escolaridade	13,858
fundamental I completo	10,660
fundamental II completo	9,899
ensino médio completo	8,068
superior completo	4,808

CONTADOR

Type	Integer
Valid	146,038
Mismatched	0
Empty	0



Minimum	35
Lower quartile	391,579
Median	761,438
Upper quartile	1,190,522
Maximum	1,556,819

TITULO_OCUPACAO

Type	String
Valid	146,038
Mismatched	0
Empty	0
Top 20 values	
ignorado	68,932
Trabalhador agropecuário em g...	6,835
Trabalhador volante da agricu...	6,494
Representante comercial autôn...	6,340
Pedreiro	6,018
Empregado doméstico nos ser...	3,517
Servente de obras	3,139
Produtor agrícola polivalente	1,999
Comerciante varejista	1,866
Motorista de carro de passeio	1,588
Caseiro (agricultura)	1,437
Vendedor de comércio varejist...	1,260
Motorista de caminhão (rotas ...	1,165
Pintor de obras	1,127
Mecânico de manutenção de aut...	983
Motofretista	882
Vigilante	868
Eletricista de instalações	635
Vendedor ambulante	607
Ajudante de motorista	551

CID10

Type	String
Valid	145,957
Mismatched	0
Empty	81
Top 20 values	
Agressão por meio de disparo ...	16,950
Lesão autoprovocada intencion...	7,365
Agressão por meio de disparo ...	6,514
Fatos ou eventos não especifi...	5,997
Agressão por meio de disparo ...	3,900
Outras quedas no mesmo nível ...	3,787
Pessoa traumatizada em um aci...	3,514
Motociclista [qualquer] traum...	3,061
Agressão por meio de objeto c...	2,781
Outras quedas no mesmo nível ...	2,427
Queda sem especificação - loc...	2,279
Ocupante [qualquer] de um aut...	2,211
Pedestre traumatizado em um a...	2,179
Agressão por meio de disparo ...	2,174
Agressão por meio de objeto c...	2,004
Afogamento e submersão em águ...	1,923
Motociclista traumatizado em ...	1,883
Intervenção legal envolvendo ...	1,834
Agressão por meio de objeto c...	1,677
Pessoa traumatizada em um aci...	1,542

UFOCOR

Type	String
Valid	146,038
Mismatched	0
Empty	0
Top 20 values	
São Paulo	22,091
Bahia	13,337
Rio de Janeiro	12,888
Minas Gerais	12,536
Pernambuco	8,700
Paraná	8,661
Ceará	8,646
Rio Grande do Sul	7,137
Pará	6,149
Goiás	6,059
Maranhão	4,880
Santa Catarina	4,406
Espírito Santo	3,481
Mato Grosso	3,185
Paraíba	2,973
Amazonas	2,738
Rio Grande do Norte	2,728
Alagoas	2,706
Piauí	2,569
Sergipe	2,004

Profile report – DOEXT21

From Cloud Storage to BigQuery – Job ID: 21535118

All Data

31 columns 149,322 rows 4 data types

- 93% valid values
- 1% mismatching values
- 6% missing values

Output destinations

Name	Location	Status
causas_externas_21	bigquery://mvp-puc-399114/obitos_datasus/causas_externas_21	-

Data sources

Name	Location	Schema validation
tabela_cid10.csv	gs://dados-complementares/tabela_cid10.csv	
tabela_ocupacoes.csv	gs://dados-complementares/tabela_ocupacoes.csv	No schema changes found
DOEXT21.csv	gs://obitos-datasus/DOEXT21.csv	No schema changes found
ibge_codigo_UF.csv	gs://dados-complementares/ibge_codigo_UF.csv	No schema changes found

Execution summary

Job Status
Completed

Job Type
Manual

Environment
Dataflow

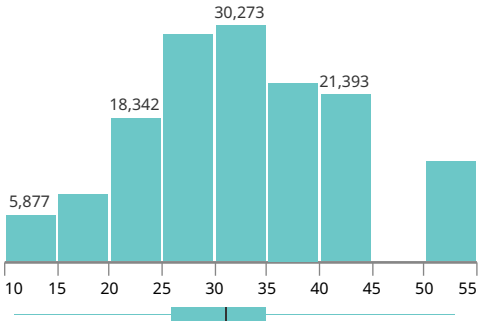
Start time
Tue, Sep 26, 2023 1:53 PM -03:00

End time
Tue, Sep 26, 2023 2:15 PM -03:00

Duration
22 min

CODUFOCOR

Type	Integer
Valid	149,322
Mismatched	0
Empty	0

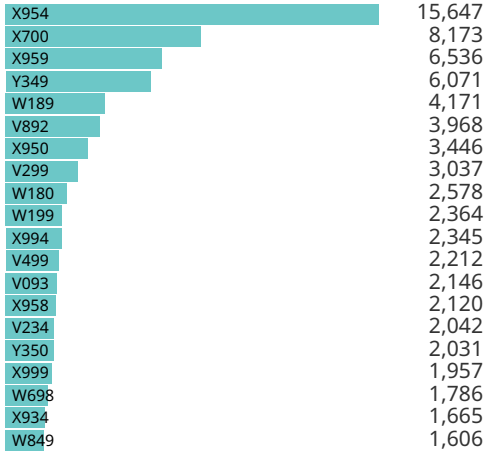


Minimum	11
Lower quartile	26
Median	31
Upper quartile	35
Maximum	53

CAUSABAS

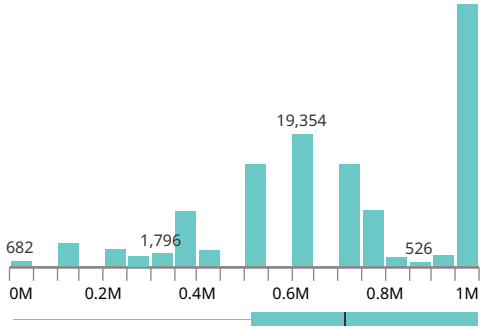
Type	String
Valid	149,322
Mismatched	0
Empty	0

Top 20 values



OCUP

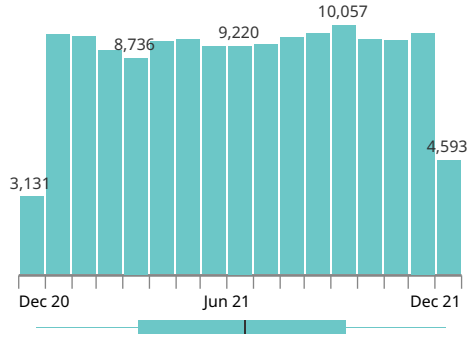
Type	Integer
Valid	119,274
Mismatched	0
Empty	30,048



Minimum	10,205
Lower quartile	518,021
Median	715,246
Upper quartile	999,991
Maximum	999,995

DTOBITO

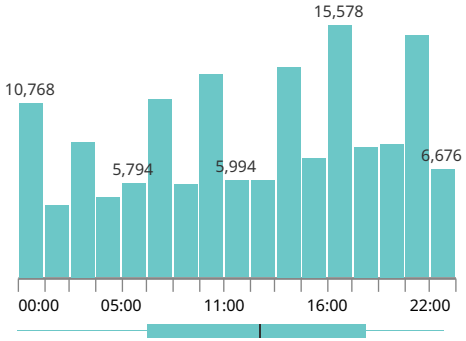
Type	Date / Time (mm*dd*yyyy)
Valid	149,322
Mismatched	0
Empty	0



Minimum	Jan 1 2021
Lower quartile	Apr 1 2021
Median	Jul 4 2021
Upper quartile	Oct 2 2021
Maximum	Dec 31 2021

HORAOBITO

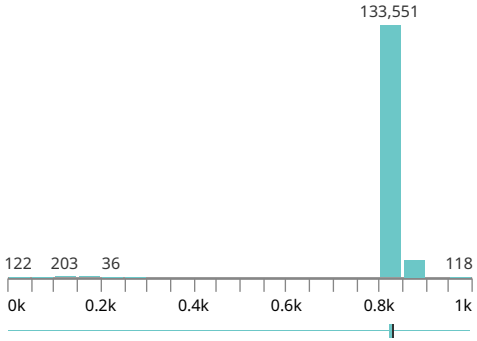
Type	Date / Time (HH:MM)
Valid	149,322
Mismatched	0
Empty	0



Minimum	00:01:00
Lower quartile	07:00:55
Median	13:05:32
Upper quartile	18:47:48
Maximum	23:01:00

NATURAL

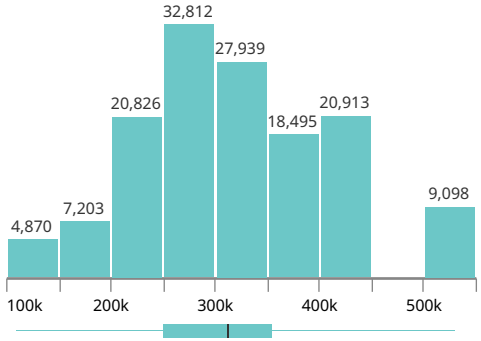
Type	Integer
Valid	143,742
Mismatched	0
Empty	5,580



Minimum	4
Lower quartile	825
Median	831
Upper quartile	835
Maximum	999

CODMUNNATU

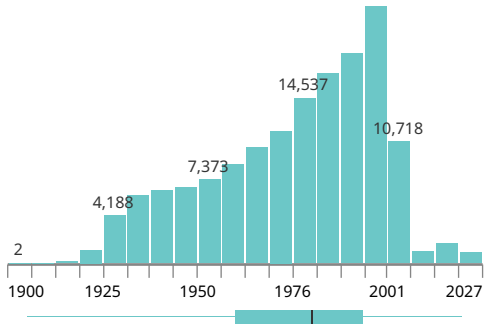
Type	Integer
Valid	142,156
Mismatched	0
Empty	7,166



Minimum	110,000
Lower quartile	251,114
Median	312,161
Upper quartile	355,000
Maximum	530,010

DTNASC

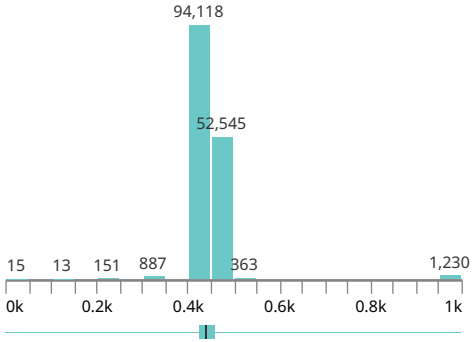
Type	Date / Time (mm*dd*yyyy)
Valid	149,322
Mismatched	0
Empty	0



Minimum	Nov 29 1905
Lower quartile	May 11 1961
Median	Sep 28 1981
Upper quartile	Apr 22 1995
Maximum	Dec 19 2021

IDADE

Type	Integer
Valid	149,322
Mismatched	0
Empty	0



Minimum	1
Lower quartile	426
Median	439
Upper quartile	460
Maximum	999

SEXO

Type	String
Valid	149,322
Mismatched	0
Empty	0

Top 3 values

masculino	118,829
feminino	30,306
ignorado	187

RACACOR

Type	String
Valid	149,322
Mismatched	0
Empty	0

Top 6 values

parda	80,572
branca	54,360
preta	10,578
ignorado	2,539
indigena	757
amarela	516

ESTCIV

Type	String
Valid	149,322
Mismatched	0
Empty	0

Top 6 values

solteiro	79,009
casado	28,272
ignorado	15,658
viuvo	11,422
separado	7,676
uniao estavel	7,285

ESC

Type	String
Valid	149,322
Mismatched	0
Empty	0

Top 6 values

de 8 a 11 anos	40,971
de 4 a 7 anos	40,884
ignorado	29,475
de 1 a 3 anos	20,688
nenhuma	9,390
12 anos e mais	7,914

ESC2010

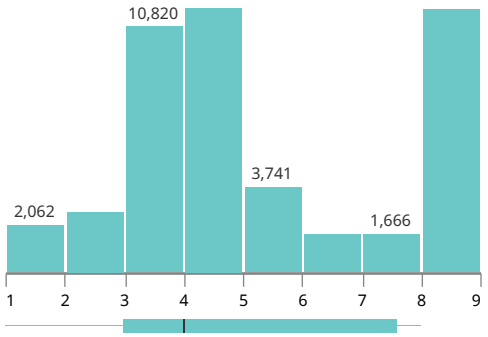
Type	String
Valid	149,322
Mismatched	0
Empty	0

Top 7 values

fundamental II (5ª a 8ª série...	40,731
fundamental I (1ª a 4ª série)	32,283
ignorado	29,733
médio (antigo 2º grau)	29,323
sem escolaridade	9,366
superior completo	5,604
superior incompleto	2,282

SERIESCFAL

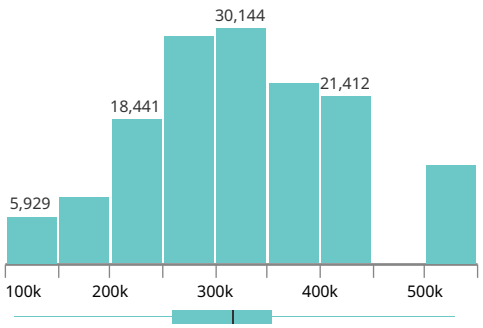
Type	Integer
Valid	45,826
Mismatched	0
Empty	103,496



Minimum	1
Lower quartile	3
Median	4
Upper quartile	8
Maximum	8

CODMUNRES

Type	Integer
Valid	149,322
Mismatched	0
Empty	0



Minimum	110,000
Lower quartile	260,138
Median	317,783
Upper quartile	355,086
Maximum	530,010

LOCOCOR

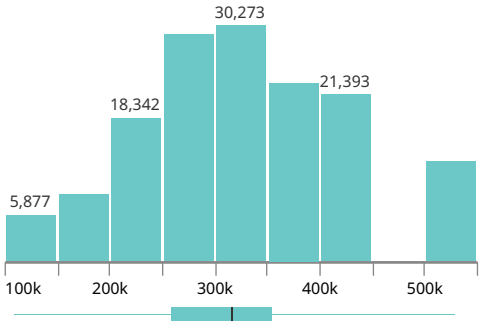
Type	String
Valid	149,322
Mismatched	0
Empty	0

Top 6 values

hospital	56,442
via pública	41,852
domicílio	24,143
outros	22,010
outros estab. de saúde	4,549
ignorado	326

CODMUNOCOR

Type	Integer
Valid	149,322
Mismatched	0
Empty	0



Minimum	110,000
Lower quartile	259,727
Median	316,585
Upper quartile	355,080
Maximum	530,010

ASSISTMED

Type	String
Valid	114,232
Mismatched	0
Empty	35,090

Top 3 values

não	58,126
sim	46,818
ignorado	9,288

NECROPSIA

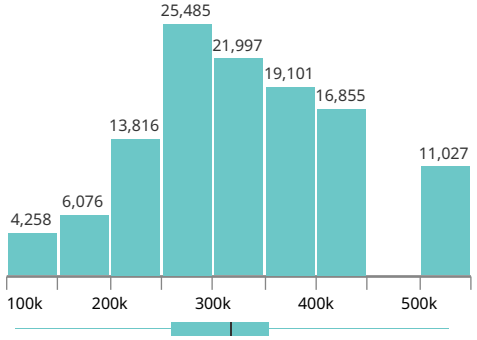
Type	String
Valid	123,502
Mismatched	0
Empty	25,820

Top 3 values

sim	95,186
não	25,770
ignorado	2,546

COMUNSVOIM

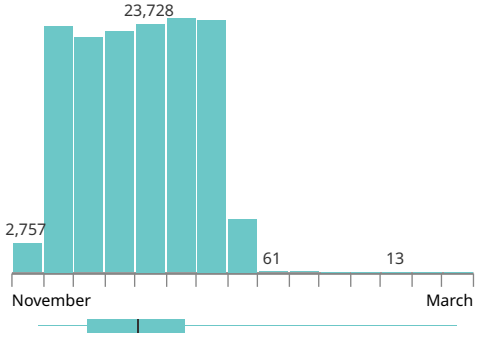
Type	Integer
Valid	118,615
Mismatched	0
Empty	30,707



Minimum	110,002
Lower quartile	261,041
Median	317,728
Upper quartile	355,189
Maximum	530,010

DTATESTADO

Type	Date / Time (mm*dd*yyyy)
Valid	149,011
Mismatched	0
Empty	311



Minimum	Jan 1 2021
Lower quartile	Apr 3 2021
Median	Jul 6 2021
Upper quartile	Oct 4 2021
Maximum	Feb 28 2023

CIRCOBITO

Type	String
Valid	148,959
Mismatched	0
Empty	363

Top 5 values

acidente	63,629
homicídio	43,384
ignorado	22,344
suicídio	14,722
outros	4,880

FONTE

Type	String
Valid	149,322
Mismatched	0
Empty	0

Top 5 values

ocorrência policial	78,823
ignorado	40,699
hospital	12,118
outra	10,076
família	7,606

TPPOS

Type	Boolean
Valid	59,974
Mismatched	42,382
Empty	46,966
Top value	
N	59,974

CAUSABAS_O

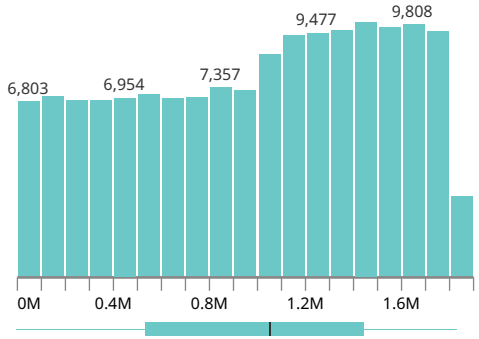
Type	String
Valid	149,234
Mismatched	0
Empty	88
Top 20 values	
X954	13,746
Y349	11,455
X700	7,420
X959	6,024
V892	4,765
X950	3,078
W189	3,025
V299	2,972
Y299	2,459
V499	2,179
R99	2,148
X994	2,069
W199	2,034
X999	1,942
X958	1,900
W180	1,820
X599	1,778
V093	1,687
W698	1,661
W849	1,562

ESCFALAGR1

Type	String
Valid	149,322
Mismatched	0
Empty	0
Top 9 values	
ignorado	29,733
fundamental II incompleto	29,146
fundamental I incompleto	20,627
ensino médio incompleto	16,796
sem escolaridade	14,599
fundamental I completo	11,656
fundamental II completo	11,585
ensino médio completo	9,576
superior completo	5,604

CONTADOR

Type	Integer
Valid	149,322
Mismatched	0
Empty	0



Minimum	35
Lower quartile	539,209
Median	1,055,623
Upper quartile	1,448,884
Maximum	1,832,645

TITULO_OCUPACAO

Type	String
Valid	149,322
Mismatched	0
Empty	0
Top 20 values	
ignorado	68,616
Representante comercial autôn...	7,149
Trabalhador agropecuário em g...	6,828
Trabalhador volante da agricu...	6,766
Pedreiro	6,069
Empregado doméstico nos ser...	3,515
Servente de obras	3,109
Comerciante varejista	2,048
Motorista de carro de passeio	1,781
Caseiro (agricultura)	1,676
Produtor agrícola polivalente	1,405
Vendedor de comércio varejist...	1,294
Motorista de caminhão (rotas ...	1,197
Pintor de obras	1,108
Motofretista	1,063
Mecânico de manutenção de aut...	1,001
Vigilante	827
Vendedor ambulante	643
Cozinheiro geral	604
Ajudante de motorista	578

CID10

Type	String
Valid	149,264
Mismatched	0
Empty	58
Top 20 values	
Agressão por meio de disparo ...	15,647
Lesão autoprovocada intencion...	8,173
Agressão por meio de disparo ...	6,536
Fatos ou eventos não especifi...	6,071
Outras quedas no mesmo nível ...	4,171
Pessoa traumatizada em um aci...	3,968
Agressão por meio de disparo ...	3,446
Motociclista [qualquer] traum...	3,037
Outras quedas no mesmo nível ...	2,578
Queda sem especificação - loc...	2,364
Agressão por meio de objeto c...	2,345
Ocupante [qualquer] de um aut...	2,212
Pedestre traumatizado em um a...	2,146
Agressão por meio de disparo ...	2,120
Motociclista traumatizado em ...	2,042
Intervenção legal envolvendo ...	2,031
Agressão por meio de objeto c...	1,957
Afogamento e submersão em águ...	1,786
Agressão por meio de disparo ...	1,665
Riscos não especificados à re...	1,606

UFOCOR

Type	String
Valid	149,322
Mismatched	0
Empty	0
Top 20 values	
São Paulo	22,908
Bahia	13,574
Minas Gerais	13,440
Rio de Janeiro	13,034
Paraná	9,186
Pernambuco	8,311
Ceará	8,128
Rio Grande do Sul	7,626
Pará	6,248
Goiás	5,928
Maranhão	4,930
Santa Catarina	4,581
Espírito Santo	3,799
Amazonas	3,304
Mato Grosso	3,268
Paraíba	2,968
Piauí	2,719
Rio Grande do Norte	2,565
Alagoas	2,483
Mato Grosso do Sul	2,118

Profile report – DOEXT22

From Cloud Storage to BigQuery – Job ID: 21535111

All Data

31 columns 152,090 rows 4 data types

- 93% valid values
- 1% mismatching values
- 6% missing values

Output destinations

Name	Location	Status
causas_externas_22	bigquery://mvp-puc-399114/obitos_datasus/causas_externas_22	-

Data sources

Name	Location	Schema validation
ibge_codigo_UF.csv	gs://dados-complementares/ibge_codigo_UF.csv	No schema changes found
DOEXT22.csv	gs://obitos-datasus/DOEXT22.csv	No schema changes found
tabela_ocupacoes.csv	gs://dados-complementares/tabela_ocupacoes.csv	No schema changes found
tabela_cid10.csv	gs://dados-complementares/tabela_cid10.csv	

Execution summary

Job Status
Completed

Job Type
Manual

Environment
Dataflow

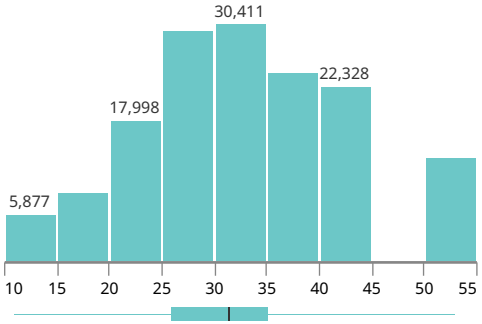
Start time
Tue, Sep 26, 2023 1:53 PM -03:00

End time
Tue, Sep 26, 2023 2:13 PM -03:00

Duration
20 min

CODUFOCOR

Type	Integer
Valid	152,090
Mismatched	0
Empty	0

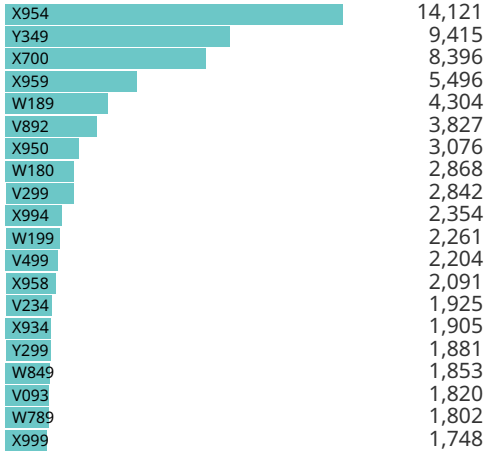


Minimum	11
Lower quartile	26
Median	31
Upper quartile	35
Maximum	53

CAUSABAS

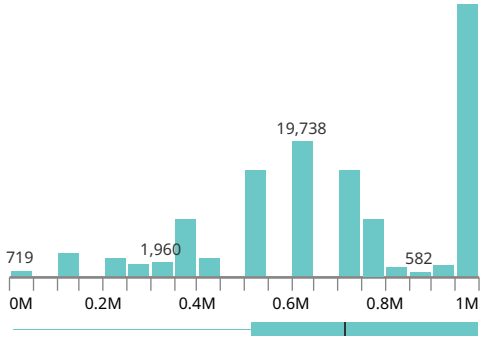
Type	String
Valid	152,090
Mismatched	0
Empty	0

Top 20 values



OCUP

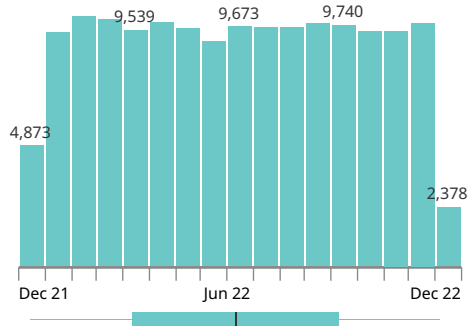
Type	Integer
Valid	123,603
Mismatched	0
Empty	28,487



Minimum	10,110
Lower quartile	517,647
Median	715,243
Upper quartile	999,991
Maximum	999,995

DTOBITO

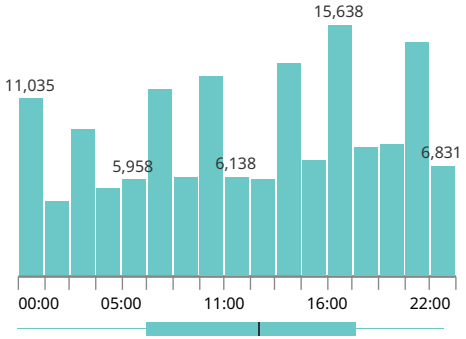
Type	Date / Time (mm*dd*yyyy)
Valid	152,090
Mismatched	0
Empty	0



Minimum	Jan 1 2022
Lower quartile	Apr 1 2022
Median	Jul 2 2022
Upper quartile	Oct 1 2022
Maximum	Dec 31 2022

HORAOBITO

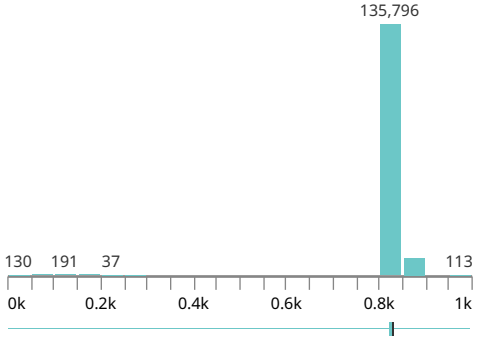
Type	Date / Time (HH:MM)
Valid	152,089
Mismatched	0
Empty	1



Minimum	00:01:00
Lower quartile	07:00:10
Median	13:01:00
Upper quartile	18:17:24
Maximum	23:01:00

NATURAL

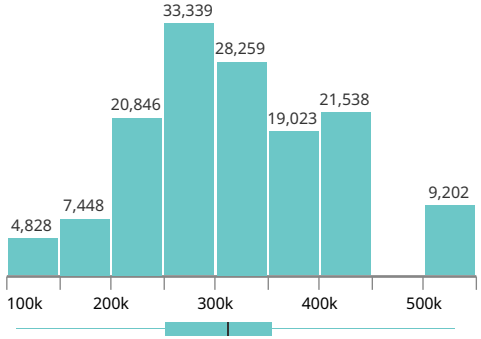
Type	Integer
Valid	146,161
Mismatched	0
Empty	5,929



Minimum	4
Lower quartile	825
Median	831
Upper quartile	835
Maximum	999

CODMUNNATU

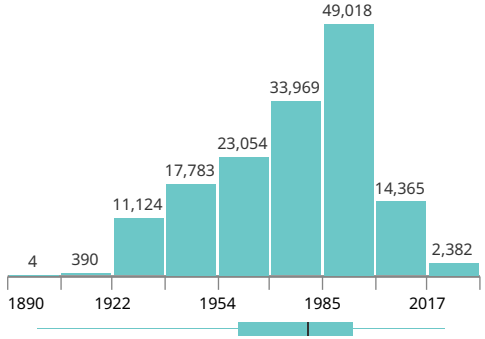
Type	Integer
Valid	144,483
Mismatched	0
Empty	7,607



Minimum	110,000
Lower quartile	253,046
Median	312,259
Upper quartile	354,996
Maximum	530,010

DTNASC

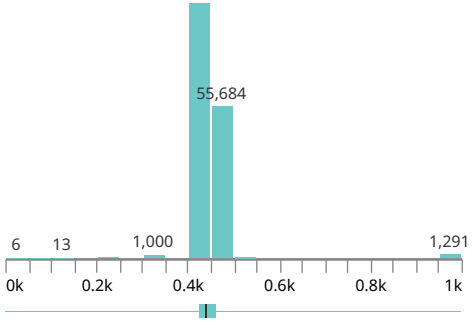
Type	Date / Time (mm*dd*yyyy)
Valid	152,089
Mismatched	0
Empty	1



Minimum	Jan 1 1900
Lower quartile	Nov 26 1960
Median	Sep 27 1981
Upper quartile	Jul 29 1995
Maximum	Dec 30 2022

IDADE

Type	Integer
Valid	152,090
Mismatched	0
Empty	0



Minimum	1
Lower quartile	427
Median	441
Upper quartile	462
Maximum	999

SEXO

Type	String
Valid	152,090
Mismatched	0
Empty	0

Top 3 values

masculino	119,971
feminino	31,952
ignorado	167

RACACOR

Type	String
Valid	152,090
Mismatched	0
Empty	0

Top 6 values

parda	80,636
branca	56,908
preta	10,915
ignorado	2,259
indigena	751
amarela	621

ESTCIV

Type	String
Valid	152,090
Mismatched	0
Empty	0

Top 6 values

solteiro	80,166
casado	28,746
ignorado	15,215
viuvo	12,362
separado	8,256
uniao estavel	7,345

ESC

Type	String
Valid	152,090
Mismatched	0
Empty	0

Top 6 values

de 8 a 11 anos	43,493
de 4 a 7 anos	40,456
ignorado	29,396
de 1 a 3 anos	20,358
nenhuma	9,701
12 anos e mais	8,686

ESC2010

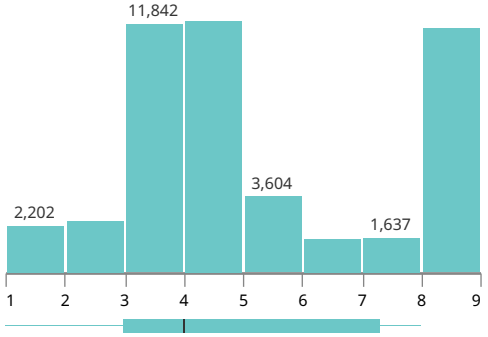
Type	String
Valid	152,090
Mismatched	0
Empty	0

Top 7 values

fundamental II (5ª a 8ª série...	39,917
fundamental I (1ª a 4ª série)	32,281
médio (antigo 2º grau)	31,696
ignorado	29,924
sem escolaridade	9,641
superior completo	6,172
superior incompleto	2,459

SERIESCFAL

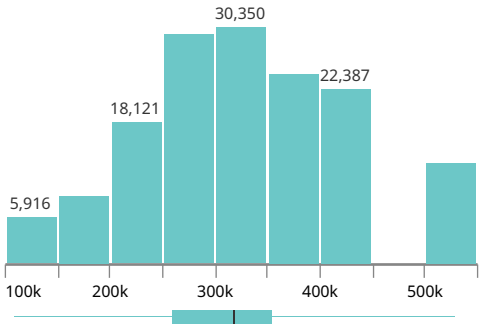
Type	Integer
Valid	46,958
Mismatched	0
Empty	105,132



Minimum	1
Lower quartile	3
Median	4
Upper quartile	7
Maximum	8

CODMUNRES

Type	Integer
Valid	152,090
Mismatched	0
Empty	0



Minimum	110,000
Lower quartile	260,461
Median	318,365
Upper quartile	355,080
Maximum	530,010

LOCOCOR

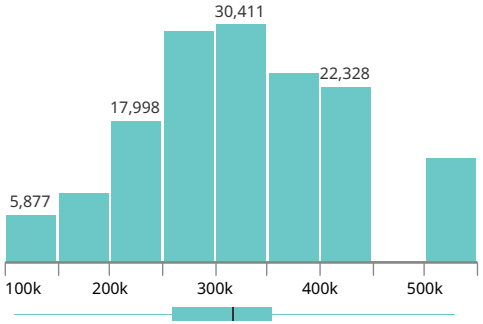
Type	String
Valid	152,090
Mismatched	0
Empty	0

Top 6 values

hospital	57,967
via pública	41,889
domicílio	24,731
outros	22,516
outros estab. de saúde	4,659
ignorado	328

CODMUNOCOR

Type	Integer
Valid	152,090
Mismatched	0
Empty	0



Minimum	110,001
Lower quartile	260,610
Median	318,004
Upper quartile	355,189
Maximum	530,010

ASSISTMED

Type	String
Valid	117,925
Mismatched	0
Empty	34,165

Top 3 values

não	59,237
sim	49,219
ignorado	9,469

NECROPSIA

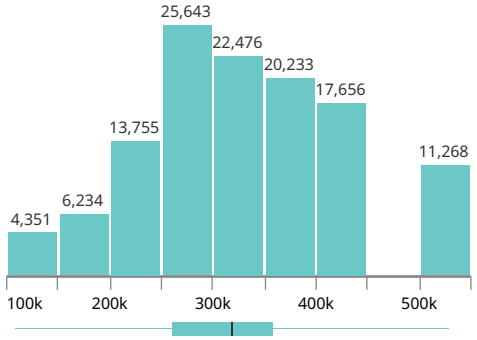
Type	String
Valid	127,101
Mismatched	0
Empty	24,989

Top 3 values

sim	100,352
não	24,111
ignorado	2,638

COMUNSVOIM

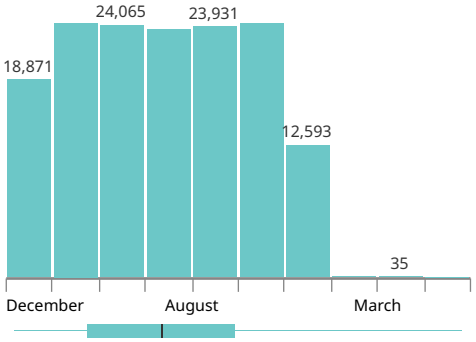
Type	Integer
Valid	121,616
Mismatched	0
Empty	30,474



Minimum	110,002
Lower quartile	261,131
Median	319,099
Upper quartile	358,924
Maximum	530,010

DTATESTADO

Type	Date / Time (mm*dd*yyyy)
Valid	151,802
Mismatched	0
Empty	288



Minimum	Jan 1 2022
Lower quartile	Apr 2 2022
Median	Jul 3 2022
Upper quartile	Oct 3 2022
Maximum	Jul 13 2023

CIRCOBITO

Type	String
Valid	151,633
Mismatched	0
Empty	457

Top 5 values

acidente	61,910
homicídio	40,342
ignorado	29,894
suicídio	14,830
outros	4,657

FONTE

Type	String
Valid	152,090
Mismatched	0
Empty	0

Top 5 values

ocorrência policial	81,655
ignorado	41,333
hospital	12,492
outra	9,509
família	7,101

TPPOS

Type	Boolean
Valid	66,784
Mismatched	33,135
Empty	52,171
Top value	
N	66,784

CAUSABAS_O

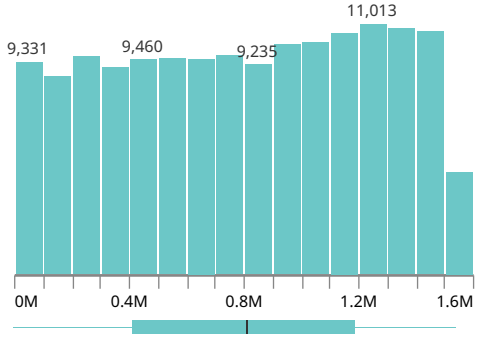
Type	String
Valid	152,014
Mismatched	0
Empty	76
Top 20 values	
X954	12,862
Y349	11,980
X700	7,873
X959	5,517
V892	4,555
W189	3,206
V299	2,823
X950	2,778
Y299	2,550
V499	2,203
X994	2,177
W180	2,118
W199	2,083
X958	1,946
X934	1,883
X599	1,861
R99	1,820
W849	1,805
X999	1,785
W789	1,716

ESCFALAGR1

Type	String
Valid	152,090
Mismatched	0
Empty	0
Top 9 values	
ignorado	29,924
fundamental II incompleto	28,276
fundamental I incompleto	20,273
ensino médio incompleto	18,136
sem escolaridade	14,962
fundamental I completo	12,008
fundamental II completo	11,641
ensino médio completo	10,698
superior completo	6,172

CONTADOR

Type	Integer
Valid	152,090
Mismatched	0
Empty	0



Minimum	2
Lower quartile	413,691
Median	812,951
Upper quartile	1,190,178
Maximum	1,542,145

TITULO_OCUPACAO

Type	String
Valid	152,090
Mismatched	0
Empty	0
Top 20 values	
ignorado	68,553
Representante comercial autôn...	7,421
Trabalhador agropecuário em g...	6,897
Trabalhador volante da agricu...	6,578
Pedreiro	6,474
Empregado doméstico nos ser...	3,776
Servente de obras	3,062
Comerciante varejista	1,920
Caseiro (agricultura)	1,798
Motorista de carro de passeio	1,764
Produtor agrícola polivalente	1,580
Vendedor de comércio varejist...	1,323
Motorista de caminhão (rotas ...	1,240
Pintor de obras	1,176
Motofretista	1,106
Mecânico de manutenção de aut...	985
Vigilante	956
Eletricista de instalações	663
Alimentador de linha de produ...	650
Vendedor ambulante	602

CID10

Type	String
Valid	152,030
Mismatched	0
Empty	60
Top 20 values	
Agressão por meio de disparo ...	14,121
Fatos ou eventos não especifi...	9,415
Lesão autoprovocada intencion...	8,396
Agressão por meio de disparo ...	5,496
Outras quedas no mesmo nível ...	4,304
Pessoa traumatizada em um aci...	3,827
Agressão por meio de disparo ...	3,076
Outras quedas no mesmo nível ...	2,868
Motociclista [qualquer] traum...	2,842
Agressão por meio de objeto c...	2,354
Queda sem especificação - loc...	2,261
Ocupante [qualquer] de um aut...	2,204
Agressão por meio de disparo ...	2,091
Motociclista traumatizado em ...	1,925
Agressão por meio de disparo ...	1,905
Contato com objeto contundent...	1,881
Riscos não especificados à re...	1,853
Pedestre traumatizado em um a...	1,820
Inalação do conteúdo gástrico...	1,802
Agressão por meio de objeto c...	1,748

UFOCOR

Type	String
Valid	152,090
Mismatched	0
Empty	0
Top 20 values	
São Paulo	24,144
Minas Gerais	13,920
Bahia	13,533
Rio de Janeiro	12,611
Paraná	9,807
Pernambuco	8,576
Ceará	8,144
Rio Grande do Sul	7,787
Pará	6,485
Goiás	5,882
Santa Catarina	4,734
Maranhão	4,676
Espírito Santo	3,880
Mato Grosso	3,523
Amazonas	3,230
Paraíba	3,014
Piauí	2,698
Rio Grande do Norte	2,480
Alagoas	2,448
Mato Grosso do Sul	2,265