Capítulo 4

Técnicas e práticas de jornalismo de dados para aquisição e gerenciamento de dados em MySQL aplicadas ao domínio da violência contra a mulher

Luciana Sá Brito, Alayne Duarte Amorim, André Viana Tardelli, Angélica Fonseca da Silva Dias, Juliana Baptista dos Santos França, Adriana Santarosa Vivacqua

Abstract

This short course aims to provide a panoramic view of the fundamental steps of acquiring and managing data practice in data journalism. It will be conducted using MySql databases resources and query language basic structures. Through the introduction of anonymization, cleansing, and transformation techniques, participants will be encouraged to practice these techniques in order to increase students' data literacy. A real case of data acquisition and management of violence against women will be applied to foment flashes of inspiration to this short course in order to amplify students' critical view and empathy about data journalism's social issues.

Resumo

Este capítulo tem como objetivo desenvolver uma visão panorâmica das etapas fundamentais de aquisição e gerenciamento de dados praticadas no âmbito do jornalismo de dados com o uso do MySql. Nele, serão introduzidos conceitos e práticas associados às técnicas de anonimização, limpeza e transformação de dados, a fim de estimular a literacia de dados do leitor. Um caso real de aquisição e gerenciamento de dados de violência contra a mulher será utilizado como objeto de aplicação das técnicas e conceitos apresentados. Espera-se que ao final do capítulo, o leitor tenha sua visão crítica e empática ampliada, acerca de temáticas sociais aderentes ao jornalismo de dados.

4.1. Introdução e Contextualização do Problema

Uma em cada 3 mulheres no mundo possui histórico de violência física e/ou sexual [World Health Organization *et al*. 2020]. Em tempos de emergência, a violência contra as mulheres continua ameaçando seriamente a saúde da população, especialmente das mulheres, sendo que o tipo mais recorrente de violência é aquela realizada pelo seu próprio parceiro íntimo [World Health Organization *et al*. 2020].

Segundo a Organização Mundial de Saúde (OMS), violência é o uso intencional de poder ou força física, em forma de ameaça ou indo às vias de fato, contra si mesmo, outra pessoa, ou contra um grupo ou comunidade, que pode resultar em ferimentos, morte, dano psicológico, desenvolvimento deficiente ou privação [World Health Organization *et al*. 2020]. Violência contra a mulher é, segundo Burelomova *et al*. (2018), o mau uso do poder pelo parceiro íntimo (homem ou mulher), que resulta em perda de dignidade, controle e segurança, bem como sentimento de impotência e aprisionamento experimentado pela mulher que é vítima direta de problemas físicos, psicológicos, abuso econômico, sexual, verbal e/ou espiritual contínuos ou repetidos. Violência contra a mulher também inclui ameaçar ou forçar mulheres o testemunho de violência por parte de seus maridos, parceiros, ex-maridos ou ex-sócios, contra seus filhos, parentes, amigos, animais de estimação e/ou bens queridos.

A Lei nº 11.340 de 2006 [Brasil 2006] nomeada Lei Maria da Penha recebeu esse nome em razão da vítima Maria da Penha Maia Fernandes (Figura 4.1) e tem como objetivo principal estipular punição adequada e coibir atos de violência doméstica e familiar contra a mulher. Segundo a referida lei, as formas de violência são classificadas como: violência física, violência psicológica, violência sexual, violência patrimonial e violência moral.

Figura 4.1. Maria da Penha, a ativista pelos direitos das mulheres que dá nome à Lei 11.340/06, ficou paraplégica por conta de agressões realizadas por seu ex marido Fonte: https://commons.wikimedia.org/wiki/File:Maria_da_Penha_em_novembro_de_2018_(cropped_2).jpg

O artigo 38 da Lei Maria da Penha prevê a criação de um Sistema Nacional de Dados e Estatísticas sobre a Violência Doméstica e Familiar contra a Mulher, que deveria ter sido implementado nos quatro anos seguintes à sua publicação pela Secretaria Especial de Políticas para as Mulheres (SPM) em articulação com outros Ministérios e órgãos da Administração Pública. Contudo, após cerca de 15 anos da publicação da Lei, ainda não é possível identificar um sistema nacional de dados sobre violência doméstica.

A ausência de uma base de dados nacional detalhada sobre a violência contra a mulher prejudica o reconhecimento da gravidade e alcance da violência doméstica no Brasil, favorecendo a manutenção das graves consequências para a saúde e segurança das mulheres do nosso país [Alves, Dumaresq e Silva 2016], problema que é agravado nas regiões do país onde há maior vulnerabilidade social [Borburema *et al*. 2017].

Este capítulo foi concebido a partir da compreensão de que é urgente aperfeiçoar as fases de coleta e tratamento de dados de violência doméstica, de forma que os resultados de sua análise reflitam essa realidade social, sendo capazes de orientar ações e políticas públicas capazes de garantir qualidade de vida e segurança às mulheres brasileiras.

4.1.1. Dados sobre Violência Contra a Mulher no Brasil

Em 2016, o Núcleo de Estudos e Pesquisas da Consultoria Legislativa do Senado nacional apontou lacunas existentes nas políticas de enfrentamento da violência contra a mulher, apontando a relevância de um banco de dados unificado sobre o tema [Alves, Dumaresq e Silva 2016]. O estudo indicou a existência de bases de dados pulverizadas, revelando a forma como os órgãos utilizam essas bases parciais para a geração de políticas através das análises desses dados.

O Observatório da Mulher Contra a Violência ─ órgão instituído pelo Senado Federal em 2016 para funcionar em conjunto com o Instituto DataSenado com a função de reunir e sistematizar as estatísticas oficiais sobre a violência contra a mulher ─ inaugurou, em março de 2019 o recurso Painel de Violência Contra Mulheres[1], que relaciona dados de diversas fontes para oferecer um panorama da violência contra a mulher no Brasil. O painel é carregado com dados coletados através do Ministério da Saúde e Conselho Nacional de Justiça, mais especificamente das fontes: Sistema de Mortalidade do Ministério da Saúde, Sistema de Informação de Agravos de Notificação do Ministério da Saúde, Pesquisa Nacional por Amostra de Domicílios do IBGE, Secretarias de Segurança Pública estaduais, Conselho Nacional de Justiça e Disque 180 da Secretaria de Políticas para Mulheres.

A navegação pelo Painel de Violência contra Mulheres contempla a visualização de dados de 2011 a 2018 através da aplicação de filtros por ocorrência de: homicídios de mulheres, notificações de saúde, boletins de ocorrência e ações do poder judiciário. Adicionalmente, podem ser aplicados filtros como ano, UF, raça, escolaridade entre outros. Contudo, ao realizar testes para a geração de visualizações de dados através do site, foram constatadas algumas inconsistências como a mostrada na Figura 4.2, que aponta a inexistência de boletins de ocorrência de violência contra mulheres no ano de 2017 ─ fonte de questionamento quanto a sua discrepância frente à realidade.

[1] https://bit.ly/3vst86w. Acesso em 27/05/2021.

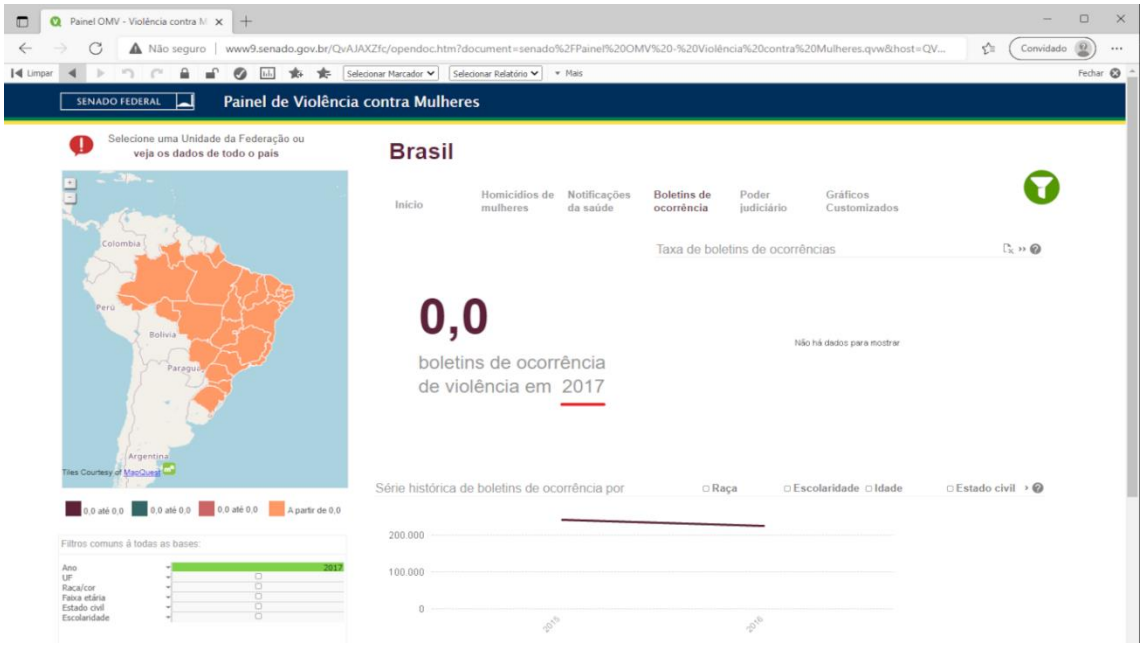

Figura 4.2. Painel Violência contra Mulher do Senado Federal. Aplicação de filtro por ano para visualizar boletins de ocorrência feitos em 2017.

O fato de não haver registros de boletins de ocorrência de violência doméstica em 2017 para todo o território nacional corrobora com a inquietação sobre a urgência da reunião dos dados como indicado na Lei Maria da Penha, porque reflete a percepção de que as bases de dados nacionais sobre violência contra a mulher realmente não estão unificadas.

Reconhecendo a pulverização das bases de dados sobre violência doméstica no Brasil, o Grupo de Pesquisa em Literacia de Dados do Programa de Pós graduação em Informática da Universidade Federal do Rio de Janeiro (PPGI - UFRJ) empreendeu um projeto com o objetivo de localizar um *dataset* referente ao registro de atendimentos a mulheres vítimas de violência doméstica e criar uma base de dados consolidada para a realização de aplicações de Ciência de Dados.

A partir dessa busca, foi realizado contato com um Centro Especializado de Atendimento à Mulher (CEAM) de um município do Estado do Rio de Janeiro, que disponibilizou um conjunto de dados para estudo. A escolha do município se deu por ser uma região historicamente marcada pela vulnerabilidade social e pela possibilidade de acesso às informações através do centro especializado.

A pesquisa realizada sobre esse conjunto de dados contribuiu para o estudo das técnicas, conceitos e práticas de jornalismo de dados abordadas neste capítulo. Ela também contribuiu para a devolução de uma base de dados consistente para o CEAM, possibilitando a visibilidade dos mais de 150 casos de violência doméstica registrados somente durante os anos iniciais da pandemia do Coronavírus (Figura 4.3), e possibilitando que outras organizações, pesquisadores e gestores produzam ou utilizem modelos para automatização da visualização dos dados para geração de conhecimento e publicações atualizadas sobre os dados de violência doméstica, fundamentando a intensificação de ações de combate a este tipo de crime.

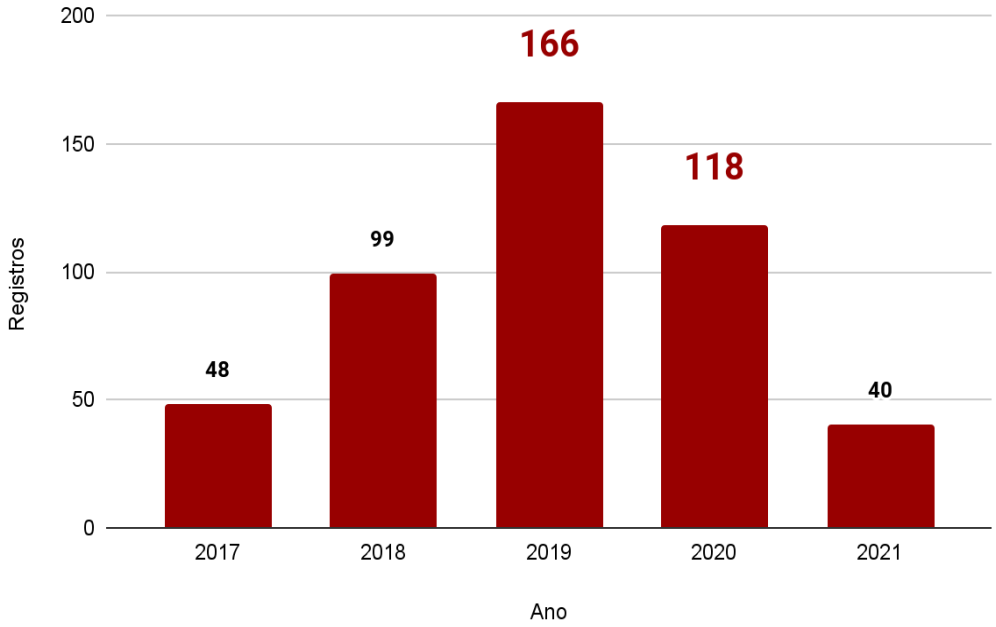

Figura 4.3. Gráfico com número de registros de violência entre os anos 2017 e 2021, com base nos dados do CEAM. A baixa quantidade de registros nos anos de 2017 e 2021 revela o início da contagem e contagem ainda em andamento.

A partir do encontro do Grupo de Literacia de dados do PPGI-UFRJ com o tema de estudo e com o conjunto de dados do CEAM, identificou-se que o Jornalismo de Dados seria uma abordagem interessante para lidar com todo o processo de descoberta de informação na base de dados. Na próxima seção, é dada uma visão geral sobre o Jornalismo de Dados.

4.1.2. Jornalismo de Dados

Compreender o jornalismo de dados pressupõe a compreensão de duas práticas jornalísticas anteriores: o jornalismo em seu sentido tradicional e o jornalismo investigativo. O jornalismo é a atividade profissional que consiste em lidar com notícias, dados factuais e divulgação de informações. É a prática de coletar, redigir, editar e publicar informações sobre eventos atuais, sendo fundamentalmente uma atividade de comunicação [Ferreira 2012]. Já o jornalismo investigativo é todo trabalho de jornalista que não se limita ao release, ao declaratório, mas vai ao cruzamento de dados e ao destrinchamento de documentos, à escuta dos personagens envolvidos e à busca de informações estatísticas [ABRAJI 2012].

O jornalismo de dados é uma evolução do jornalismo investigativo [Appelgren e Nygren 2014] que pode se apresentar como uma mistura entre o jornalismo tradicional, o jornalismo investigativo e um jornalismo que usa bases de dados de grandes volumes (*big data*), com o auxílio de mecanismos eletrônicos para a análise e o processamento de dados, a fim de encontrar e comunicar verdades sociais importantes. O que torna o jornalismo de dados diferente dos demais, é a capacidade de combinar o "faro" tradicional para notícias com a capacidade de contar histórias convincentes por meio de dados [Bounegru *et al*. 2012].

Os produtos do jornalismo de dados são, entre outros:

- conjuntos de dados abertos (ou *datasets*);
- infográficos;
- *apps* de notícias;
- *blogs* de dados;
- *sites* de dados;
- visualizações interativas.

Alguns exemplos de visualizações interativas são:

- Projeto "Sobreviventes", que investiga a violência de gênero na Colômbia - http://especiales.datasketch.co/sobrevivientes/index.html;
- Projeto "Lost Mothers", que dá visibilidade às altas taxas de mortalidade materna nos Estados Unidos - https://www.propublica.org/article/lost-mothers-maternal-health-died-childbirth-pregnancy;
- Projeto News 18 sobre a falta de reconhecimento do trabalho realizado pelas camponesas na índia - https://www.news18.com/news/immersive/women-farmers-of-india.html.

Neste capítulo não foi utilizado um *dataset* caracterizado como *big data*, porque a finalidade pedagógica de interesse é prover conhecimentos básicos de gerenciamento de dados em MySQL. Conhecimentos esses, que uma vez aprendidos e praticados com um *dataset* de pequenas proporções, poderá ser aplicado em conjuntos de dados mais volumosos.

4.1.3. Tipos de Acesso, Licença e Formato

Os dados que o jornalista utilizará na sua investigação poderão ser fornecidos por entidades públicas por meio de licenças mais restritivas (fechadas) ou mais permissivas (abertas). Há várias questões jurídicas presentes nessa questão, que variam em função da licença adotada por cada entidade.

As licenças abertas, em geral, permitem o livre uso e reuso dos dados [Villars *et al*. 2011]. Como exemplo de disponibilização de dados abertos, temos o *site* atual do Ministério da Saúde[2], que informa em seu rodapé: "Todo o conteúdo deste site está publicado sob a licença Creative Commons Atribuição - SemDerivações 3.0 Não Adaptada". Já como exemplo de disponibilização de dados fechados, temos o site antigo do Ministério da Saúde[3], que informa em seu rodapé "Copyright © Ministério da Saúde. Todos os direitos reservados 2013/2021".

Os governos em geral mantêm plataformas com dados em acesso aberto (ou livre), que são os dados que podem ser acessados gratuitamente mesmo que a liberdade de uso desses dados possa variar segundo sua licença. Alguns *sites* de dados de livre acesso governamentais são:

[2] https://www.gov.br/saude/pt-br. Disponibilizado em agosto de 2021.
[3] https://antigo.saude.gov.br/. Disponibilizado em agosto de 2021.

- Governo da Colômbia: https://www.datos.gov.co/;
- Governo da Argentina: https://www.datos.gob.ar/;
- Governo do Brasil: https://dados.gov.br/;
- Governo do México: https://datos.gob.mx/;
- Governo dos EUA: https://catalog.data.gov/dataset;
- Governo da Rússia: https://data.gov.ru/?language=en.

É importante não confundir acesso aberto com licença aberta. Dados de acesso aberto são os dados que se pode acessar gratuitamente, enquanto os dados de licença aberta, mais conhecidos como os "dados abertos" são aqueles dados que podem ser consumidos livremente e redistribuídos em sua versão original ou modificada.

A Open Knowledge Foundation define que "Dados e conteúdo abertos podem ser usados, modificados e compartilhados livremente por qualquer pessoa para qualquer propósito" sendo relevante observar a licença desses dados, pois requisitos de preservação de procedência e abertura podem existir [Open Knowledge Foundation 2021]. A partir da tomada de consciência do valor social dos dados, especialistas em dados abertos têm discutido sobre a abertura de dados do setor privado para ações que beneficiem o interesse público, fomentando a colaboração por meio da troca de dados entre os setores (público/privado), o chamado *Data Collaboratives* [GovLab, 2021].

Além do acesso e da licença, é necessário observar o formato no qual os dados estão apresentados. Formatos abertos de arquivos, em geral, podem ser processados por diversos programas de computador e incluem formatos como .jpg, .mp3, .pdf, bem como conjuntos de formatos, tais como os *OpenDocuments* (.odt, .ods, etc.) e os *Office Open XML* (.docx, .pptx, .xlsx, etc.). Os formatos fechados são protegidos por patentes e costumam ser lidos apenas pelos programas oficiais das empresas que detêm essas patentes. Como exemplos de formatos fechados, temos os .xls, .doc, .ppt, .bmp e outros. O formato de arquivo mais adequado para a publicação de dados abertos é o .csv (*Comma Separated Values* - Valores Separados por Vírgulas), que apresenta os valores organizados como mostra a Figura 4.4.

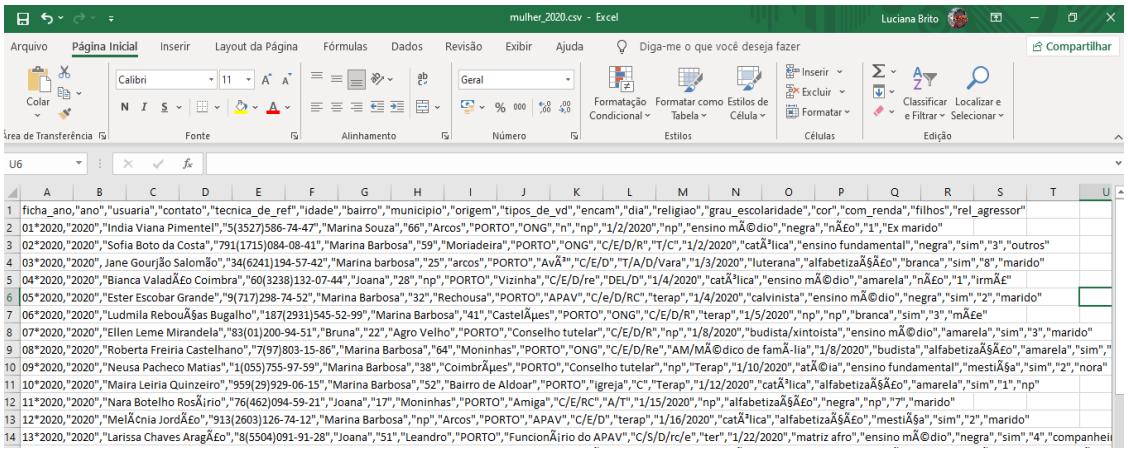

Figura 4.4. Tabela de dados em formato .csv exibidos no Microsoft Excel. © Microsoft Corporation.

4.2. Coleta dos Dados

O *dataset* é uma parte essencial de um projeto de descoberta de informações em uma base de dados, e encontrar um conjunto de dados que atenda às necessidades de pesquisa é boa parte do trabalho. De uma forma geral, os dados podem ser obtidos em repositórios disponibilizados para acesso e uso público (veja exemplo no Vídeo 1[4]) ou por iniciativas próprias de coleta, através de metodologia e recursos específicos para esse fim, como: pesquisas, questionários, entrevistas e outras formas que permitam reunir informações úteis e organizadas no formato desejável.

Ferramentas como o Kaggle[5] facilitam a aquisição de dados através dos *datasets* disponibilizados na sua plataforma, enquanto que ferramentas como o Google Data Search[6] facilita a aquisição de dados que estão disponibilizados em diferentes plataformas. O Kaggle é um dos ambientes mais conhecidos para competições de Ciência de Dados, armazenando e disponibilizando dados sobre assuntos diversos. O Google Data Search é uma ferramenta de busca que localiza um *dataset* onde quer que ele esteja hospedado. Seu funcionamento é semelhante ao mecanismo de busca do Google Scholar.

Contudo, apesar de podermos contar com diversos meios eletrônicos de disponibilização de conjuntos de dados, é possível não encontrar um conjunto que atenda às expectativas, ou ainda, não encontrar um que esteja disponibilizado para acesso público. Essa última foi a situação que o grupo de pesquisa se deparou quando iniciou a investigação sobre violência doméstica. Foram realizadas buscas por bases de dados abertos governamentais relacionados ao tema de pesquisa, considerando o período dos últimos 10 anos que, em teoria, corresponde ao tempo máximo indicado na Lei Maria da Penha para a implantação do Sistema Nacional de Dados e Estatísticas sobre a Violência Doméstica e Familiar contra a Mulher.

A partir da constatação da dificuldade de encontrar uma base de dados unificada e com acesso público sobre o tema, este estudo empenhou esforços em localizar os dados referentes aos registros de violência doméstica contra mulher em um dos municípios do Estado do Rio de Janeiro, conforme descrito na seção 4.1.1. A primeira forma de busca aconteceu através de contato telefônico com a Secretaria de Assistência Social do município, que direcionou os pesquisadores para o Centro Especializado de Atendimento à Mulher (CEAM).

Em entrevista com os especialistas do CEAM, foi identificado que os registros de atendimentos a mulheres vítimas de violência doméstica acontecem em formulários impressos em papel, sendo posteriormente digitados em planilhas eletrônicas. Constatou-se que a geração dessas planilhas faz parte de uma iniciativa independente do CEAM, uma vez que não há uma indicação de que esses dados alimentam uma base integrada no município ou em outras instâncias governamentais. Contudo, os especialistas relataram que remetem relatórios periódicos aos gestores do município a partir desses dados planilhados.

O *dataset* do CEAM contém os registros dos atendimentos realizados na unidade às mulheres vítimas de violência doméstica e está dividido em 5 planilhas eletrônicas que refletem as características da população atendida, bem como dados contextuais sobre a

[4] O Vídeo 1 pode ser acessado em https://youtu.be/EvYvKUeXqis.
[5] https://www.kaggle.com/.
[6] https://datasetsearch.research.google.com/.

violência sofrida no período de outubro de 2017 a março de 2021, época da cessão do conjunto de dados pela diretoria do CEAM. Devido à natureza sensível dos dados, o CEAM não os disponibiliza como dados abertos nem autoriza seu acesso ao público. Por esse motivo, foi requerido aos integrantes do grupo de pesquisa um termo de uso de dados e anonimato dos indivíduos listados nos registros. Esta pesquisa se reserva na obrigação de não publicar a base de dados e seu conteúdo.

Para a realização das atividades práticas descritas neste capítulo será utilizado um *dataset* fictício, chamado *mulher_2020*[7], criado a partir do conjunto de dados da CEAM. Na próxima seção, será abordado o processo de criação de uma base de dados utilizando o *dataset mulher_2020*, à semelhança da criação da base de dados gerada com o *dataset* CEAM através dos processos de carregamento, anonimização, limpeza e transformação.

4.3. Criação da Base de Dados

Um banco de dados é uma coleção organizada de dados e informações estruturadas, geralmente armazenadas eletronicamente em um sistema de computador [Oracle 2021]. Os bancos de dados têm como objetivo dar suporte aos sistemas de informação e geralmente são controlados por um Sistema de Gerenciamento de Banco de Dados (SGBD), que gerencia a manipulação dos dados, dando suporte a consultas e permitindo criar cópias de segurança, entre outras diversas tarefas [Bazzi 2013].

O MySQL é um SGBD da categoria dos *softwares livres*, pertencente à Oracle Corporation, que usa como interface a Linguagem de Consulta Estruturada (SQL). Algumas grandes corporações utilizam o MySQL para lidar com grandes volumes de dados, como o Facebook, a Google e a Adobe [Oracle Corporation and/or Its Affiliates 2021].

Este capítulo utiliza o MySQL como suporte para a criação de uma base de dados utilizando o conjunto de dados *mulher_2020*. Através das consultas SQL serão realizadas as fases de anonimização, limpeza e transformação dos dados a fim de gerar uma base de dados consolidada para a realização de análises que permitam mostrar para a sociedade características relevantes da violência contra mulheres em uma determinada localidade.

Na próxima seção, será dado início às atividades práticas de jornalismo de dados, iniciando com a ativação do banco de dados MySQL.

4.3.1. Banco de Dados MySQL em Ação

Para iniciar a criação da base de dados para análise dos dados de violência contra a mulher, será configurado um banco de dados MySQL. Para realizar essa configuração, primeiramente é necessário ativar o servidor MySQL e depois determinar a escolha de uma *interface* gráfica para o gerenciamento dos dados da base.

Para iniciar o servidor de banco de dados, será utilizada a ferramenta XAMPP, que é uma distribuição Apache e constitui-se como um pacote que contém os principais servidores de código aberto existentes, entre eles o MySQL. É possível obter a ferramenta

[7] O *dataset* mulher_2020 pode ser acessado no repositório https://github.com/Lu-Brito/Escola-Regional-de-Sistemas-de-Informa-o-2021---UNIRIO.

XAMPP através do *site* XAMPP *Installers and Downloads for Apache Friends*[8]. Após baixá-lo, você deverá abrir o arquivo executável e realizar o passo a passo de instalação que você encontrará no Vídeo 2[9].

Depois de instalado o servidor, deverá ser instalado adicionalmente o *software* que permitirá a criação da base de dados e realização das consultas SQL. Para esta prática, será utilizado o MySQL Workbench, que é a ferramenta de *interface* gráfica (GUI) oficial para o MySQL. Ela possibilitará a criação de um banco de dados, o carregamento dos dados que serão trabalhados e a realização de consultas nos dados estruturados.

É possível obter a ferramenta MySQL Workbench através do *site* Download MySQL Workbench. Após baixá-lo, será possível abrir o arquivo executável e realizar o passo a passo de instalação que você encontrará no Vídeo 3[10].

Instalados os programas, para ativar o servidor, basta abrir o XAMPP e ativar o servidor MySQL clicando no botão *Start*, como mostra a Figura 4.5.

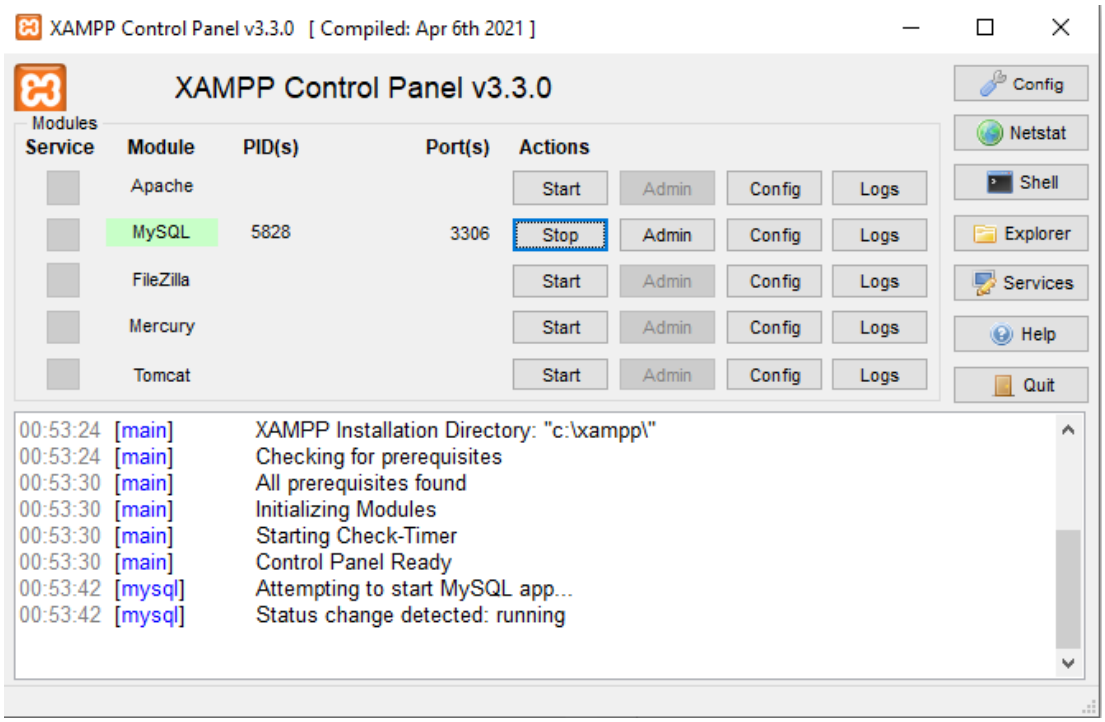

Figura 4.5. Para ativar o banco de dados MySQL, basta clicar em "*Start*".

Depois de ativar o servidor, será o momento de criar o banco de dados para que as consultas possam ser realizadas. Todo o processo de criação do banco de dados e *upload* da tabela *mulher_2020* poderá ser acompanhado no Vídeo 4[11]. Após isso, será possível visualizar a tabela, com as suas colunas no MySQL Workbench, tal como na Figura 4.6.

[8] https://www.apachefriends.org/pt_br/index.html. Acesso em outubro de 2021.
[9] O Vídeo 2 pode ser acessado em: https://youtu.be/id_DtSCfJUg.
[10] O Vídeo 3 pode ser acessado em https://youtu.be/OjS9mwNskc0.
[11] O Vídeo 4 pode ser acessado em https://youtu.be/70H4XcFVQI0.

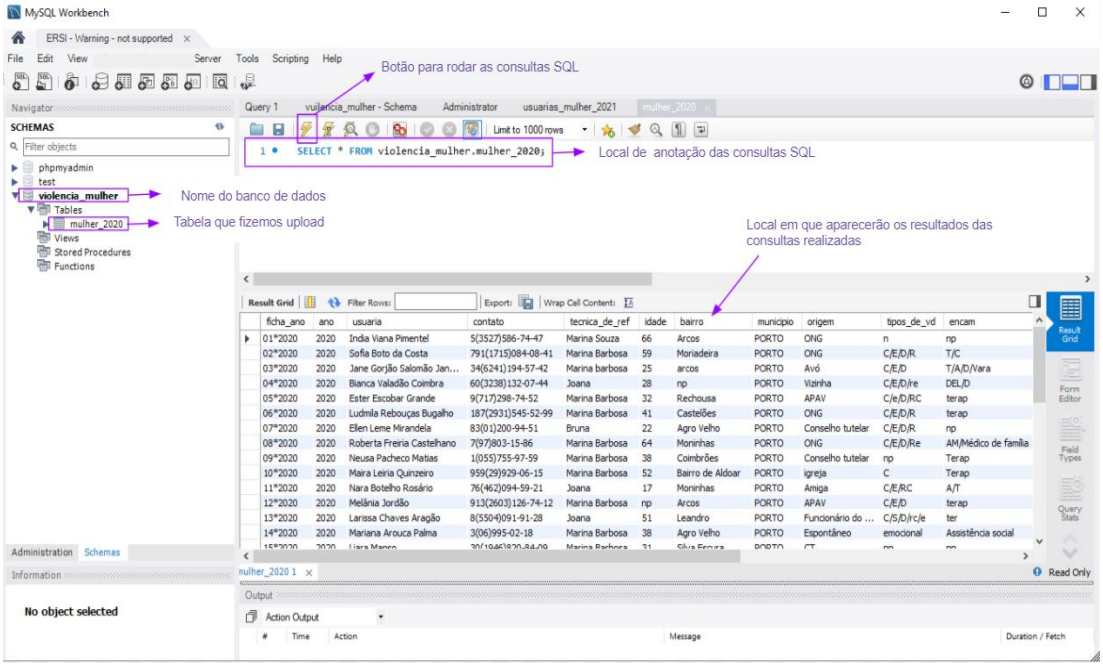

Figura 4.6. Visão geral da disposição dos elementos do MySQL Workbench que serão mais utilizados durante a realização das consultas SQL propostas neste capítulo.

Após a montagem do banco, será o momento de realizar a anonimização dos dados. Na seção 4.4 serão apresentadas algumas técnicas de anonimização de dados, além de atividades práticas para esta finalidade.

4.4. Anonimização dos Dados

A Lei 13.709 de 2018, conhecida como Lei Geral de Proteção de Dados (LGPD), prevê o tratamento dos dados pessoais de pessoa física ou jurídica de direito público ou privado para a preservação dos direitos fundamentais de liberdade, privacidade e o livre desenvolvimento da personalidade. Entre as normas gerais contidas na LGPD, a anonimização de dados que permitam identificar direta ou indiretamente um indivíduo, tem como objetivo a preservação do direito da pessoa.

Segundo a LGPD, a anonimização trata da "utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo". Durante a análise de uma base de dados é possível haver dados pessoais que identifiquem um indivíduo, sem que essa vinculação seja desejada. Para fins de preservação, esses dados devem ser anonimizados [Brasil 2018] e, para isso, devem ser aplicadas técnicas específicas para a desvinculação dos dados sem que seja prejudicada a possibilidade de utilização da base de dados.

4.4.1. Técnicas de Anonimização

Anônimo é aquele que não apresenta nome ou assinatura, indivíduo desconhecido [Michaelis 2021]. A definição do conceito de dados pessoais pode seguir uma orientação expansionista (a partir da delimitação de "pessoa identificável") ou reducionista ("pessoa identificada"), respectivamente alargando ou restringindo o escopo de aplicação da LGPD [Bioni 2019]. A princípio, o processo de anonimização pode estar atrelado aos dados pessoais numa orientação reducionista, contudo, é necessária uma análise contextual para verificar se na combinação de outros dados, como num quebra-cabeça, a identificação da pessoa seria possível, segundo a orientação expansionista.

Neste sentido, há técnicas que buscam eliminar elementos identificadores de uma base de dados. Bioni (2019) esclarece que o processo de anonimizar um dado é composto por técnicas que buscam eliminar elementos identificadores ou identificáveis da base em que estão organizados. O autor aponta quatro técnicas para nortear o processo: (i) a supressão; (ii) generalização; (iii) randomização e (iv) pseudoanonimização. A escolha da técnica ou da combinação delas precisa considerar os dados que devem ter seus vínculos quebrados com seus respectivos titulares.

4.4.1.1. Supressão

A técnica de supressão utiliza como recurso para a anonimização, a exclusão de campos da tabela de dados ou a substituição de parte de caracteres do campo de identificação. Em alguns casos, a substituição de caracteres pode não ser suficiente para anonimizar a base. É o caso do CPF, que é um campo que trata da identificação exclusiva de um indivíduo. Nesta situação, vale a pena considerar a exclusão do campo.

4.4.1.2. Generalização

A generalização propõe a substituição de um dado por outro que traz significado mais geral. Em algumas situações pode-se considerar a supressão como uma forma de generalização. Um exemplo é a supressão do CEP. Os últimos dígitos dos dados de CEP de uma base podem ser excluídos da mesma e ainda assim isso não ser suficiente para gerar a sua anonimização. Isto porque o CEP pode ser combinado com outros dados que fazem parte da base. Assim, a generalização propõe a substituição de um dado por outro que traga um significado mais geral, como por exemplo, a substituição do CEP pelo bairro ou a substituição da data de nascimento pela faixa etária.

4.4.1.3. Randomização

A randomização é uma técnica que busca mascarar uma informação, misturando os valores da tabela, sem afetar as análises, apenas visando não identificar seus titulares. Considerando uma tabela com a abordagem relacional, os dados de uma coluna (um campo) são alternados entre suas linhas em suas ordens, sem alterar seus valores. Por exemplo, considere a Tabela 4.1 sendo uma tabela original de dados.

Tabela 4.1. Tabela Original de Dados

Data de Nascimento	Código postal	Cor	Violência
27/05/2000	4618-4775	Preta	Psicológica
30/09/2001	8798-8956	Parda	Física
28/03/1960	4618-1228	Parda	Psicológica

Na randomização, os dados são alternados entre as linhas de registros. A Tabela 4.2 ilustra um exemplo da tabela após a randomização dos dados.

Tabela 4.2. Tabela com Dados Randomizados

Data de Nascimento	Código postal	Cor	Violência
27/05/2000	8798-8956	Parda	Psicológica
30/09/2001	4618-1228	Parda	Psicológica
28/03/1960	4618-4775	Preta	Física

É importante ter cuidado na aplicação desta técnica, pois ela pode causar impactos na análise dos dados. As respostas para perguntas que dependem da combinação entre dados de duas ou mais colunas podem ser comprometidas.

4.4.1.4. Pseudoanonimização

A pseudoanonimização, segundo a LGPD, diz respeito ao "tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão pelo uso de informação adicional mantida separadamente pelo controlador em ambiente controlado e seguro". Nessa técnica, a base de dados é dividida em duas partes, a primeira com os dados genéricos e a segunda com os dados sensíveis.

É importante notar que não há uma técnica melhor ou combinação perfeita para direcionar o processo de anonimização, sendo essencial analisar cada contexto em que o tratamento será implementado, para que os indivíduos titulares dos dados anonimizados não sejam reindentificados.

4.4.2. Anonimizando a Base através da Pseudoanonimização

A fase de anonimização da base de dados, dependerá da forma de anonimização escolhida. Para a realização da anonimização dos dados do *dataset* CEAM, foi feita a divisão do conjunto de dados em duas tabelas (*cadastro_ceam_usuarias* e *cadastro_ceam_atual*), em um banco de dados MySql. Para isto foi criada uma coluna com um identificador único, a fim de relacionar as tabelas.

Na tabela *cadastro_ceam_usuarias* foram reunidas 5 colunas, sendo 3 delas relativas aos dados que representam as informações julgadas como as mais sensíveis da base: nome, telefone e idade da usuária. Na tabela *cadastro_ceam_atual* foram reunidas as demais colunas constantes do *dataset* original, de modo que qualquer tarefa de mineração de dados pode ser realizada sem a necessidade de compartilhamento dos dados pessoais das usuárias. As estruturas finais das tabelas *cadastro_ceam_usuarias* e *cadastro_ceam_atual*, na base de dados já consolidada, podem ser vistas na Tabela 4.3 e na Tabela 4.4, que seguem.

Tabela 4.3. Estrutura de dados da tabela *cadastro_ceam_usuarias* com os dados pessoais das usuárias.

Name	Type	Length	Decimals	Allow Null
id	int	11	0	
ficha_ano	varchar	255	0	x
ano	varchar	255	0	x
usuaria	varchar	255	0	x
contato	varchar	255	0	x
idade	varchar	255	0	x

Tabela 4.4. Estrutura de dados da tabela *cadastro_ceam_atual* com os dados efetivamente utilizados na análise.

Name	Type	Length	Decimals	Allow Null
id	int	11	0	
ficha_ano	varchar	255	0	x
ano	varchar	255	0	x
tecnica_de_ref	varchar	255	0	x
idade	varchar	255	0	x
bairro	varchar	255	0	x
município	varchar	255	0	x
distrito	varchar	255	0	x
origem	varchar	255	0	x
tipos_de_vd	varchar	255	0	x
encam	varchar	255	0	x
dia	varchar	255	0	x
religiao_SN	varchar	255	0	x

grau_escolaridade	varchar	255	0	x
cor	varchar	255	0	x
com_renda	varchar	255	0	x
filhos	varchar	255	0	x

Para a realização da atividade proposta neste capítulo, escolheu-se criar uma tabela com as colunas que possuíam os dados pessoais das vítimas e outra tabela com os demais dados. A forma escolhida para realizar o relacionamento entre as tabelas foi através da criação de uma coluna com um identificador único (um número único para cada vítima), representada pela coluna <id>.

De forma prática, as operações que foram realizadas no banco seguiram o passo a passo descrito adiante.

4.4.2.1. Criação da coluna id na tabela *usuarias_mulher_2020*

A criação da coluna <id> é o primeiro passo importante no processo de anonimização da base, pois através da coluna <id> se pode criar um relacionamento entre a tabela com os dados pessoais das vítimas e a outra tabela, com os demais dados. Para criar a coluna <id>, será realizada a seguinte instrução SQL:

```
ALTER TABLE mulher_2020 ADD id INT NOT NULL PRIMARY KEY auto_increment;
```

Esta instrução criará a coluna <id> na tabela. Quando terminar a execução, será possível localizar a coluna <id>, que será criada após a última coluna. É possível clicar em cima do cabeçalho da coluna e arrastá-la para a esquerda, posicionando-a como a primeira coluna da tabela. É uma convenção utilizar a coluna <id> como a primeira coluna em tabelas desta finalidade, conforme mostra a Figura 4.7.

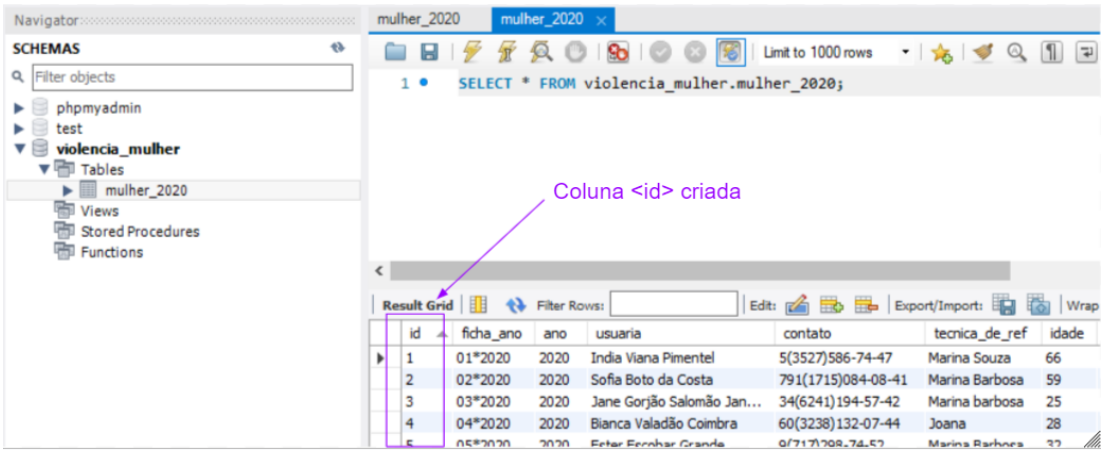

Figura 4.7. Resultado da criação da coluna <id>.

Agora que a coluna <id> está presente na tabela, é possível replicá-la a fim de criar uma tabela somente com os dados pessoais das vítimas e outra com os demais dados.

4.4.2.2. Criação da tabela *usuarias_mulher_2020* a partir da duplicação da tabela *mulher_2020* original do *dataset*

Para criar tabela *usuarias_mulher_2020* foi realizada a duplicação da tabela *mulher_2020* com a instrução SQL:

```
CREATE TABLE usuarias_mulher_2020 SELECT * FROM mulher_2020;
```

Esta instrução seleciona todas as colunas da tabela *mulher_2020* e cria uma tabela idêntica no banco *violencia_mulher*, nomeando-a como *usuarias_mulher_2020*. A nova tabela criada no banco aparecerá logo abaixo da tabela *mulher_2020*, conforme a Figura 4.8.

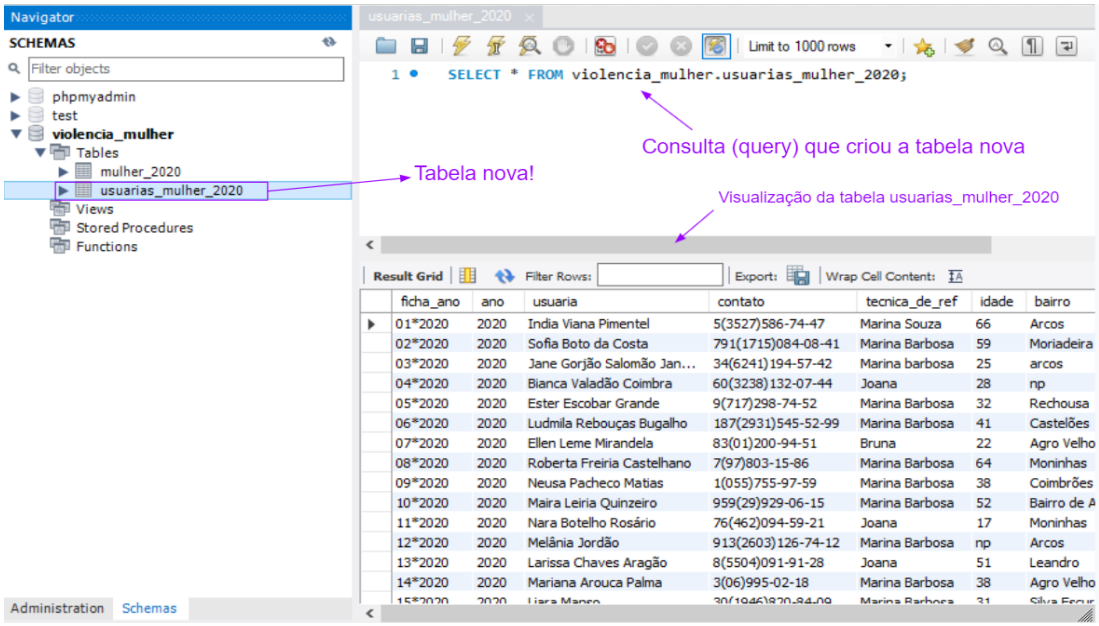

Figura 4.8. Resultado da duplicação da tabela *mulher_2020*.

Seguem também algumas recomendações em casos de erro:

1. Caso a tabela *usuarias_mulher_2020* não apareça, talvez os dados não tenham sido atualizados automaticamente. Para isso, clique com o botão direito do mouse no banco <*violencia_mulher*> e selecione a opção <Refresh All>. Faça o mesmo em <Tables>, caso ainda assim a tabela nova não apareça.
2. Para visualizar a tabela *usuarias_mulher_2020*, basta clicar no ícone em formato de tabela, que se encontra ao lado do nome da tabela, tal qual está indicado na Figura 4.9.

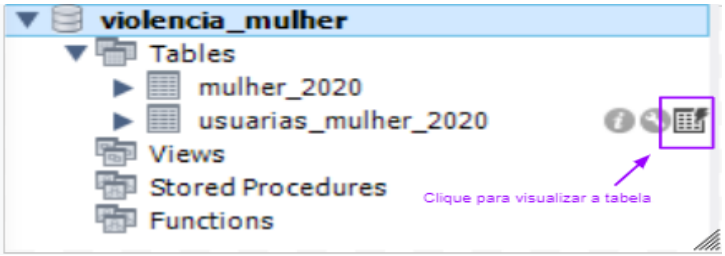

Figura 4.9. Ícone para visualização da tabela.

4.4.2.3. Organização da tabela mulher_2020 através da supressão das colunas da tabela *usuárias_mulher_2020*, exceto as colunas <id>, <ficha_ano> e <ano>

A supressão das colunas com os dados pessoais da tabela *mulher_2020* foi realizada considerando que os dados pessoais eram os das colunas <usuaria>, <contato>, <idade> e que as colunas <id>, <ficha_ano> e <ano> deveriam permanecer tanto na tabela *mulher_2020* quanto na tabela *usuarias_mulher_2020*. As colunas suprimidas a fim de anonimizar a tabela mulher_2020 foram <usuaria>, <contato> e <idade>. Para a supressão de colunas em uma tabela pode ser usada a estrutura:

```
ALTER TABLE nome da tabela DROP COLUMN coluna a ser suprimida;
```

Esta instrução SQL altera a tabela *mulher_2020*, retirando as colunas indicadas. Desta forma, foi realizada a instrução:

```
ALTER TABLE mulher_2020
      DROP COLUMN usuaria,
      DROP COLUMN contato,
      DROP COLUMN idade;
```

Após a supressão das colunas com os dados pessoais da tabela *mulher_2020*, as colunas que não foram consideradas dados pessoais na tabela *usuarias_mulher_2020* também foram suprimidas.

4.4.2.4. Organização da tabela *usuarias_mulher_2020* através da supressão das colunas com os dados que não foram considerados dados pessoais

As colunas suprimidas foram: <tecnica_de_ref>, <bairro>, <municipio>, <origem>, <tipos_de_vd>, <encam>, <dia>, <religiao>, <grau_escolaridade>, <cor>, <com_renda>, <filhos> e <rel_agressor>. A supressão das colunas foi feita através da consulta:

```
ALTER TABLE usuarias_mulher_2020
      DROP COLUMN tecnica_de_ref,
      DROP COLUMN bairro,
      DROP COLUMN municipio,
      DROP COLUMN origem,
      DROP COLUMN tipos_de_vd,
      DROP COLUMN encam,
      DROP COLUMN dia,
      DROP COLUMN religiao,
      DROP COLUMN grau_escolaridade,
      DROP COLUMN cor,
      DROP COLUMN com_renda,
      DROP COLUMN filhos,
      DROP COLUMN rel_agressor;
```

A partir da supressão das colunas desejadas, obteve-se então a consolidação da tabela *usuarias_mulher_2020*, que figura na base de dados com o objetivo de identificar as usuárias somente caso necessário, viabilizando o sigilo dos dados pessoais das vítimas durante as análises de dados realizadas com a tabela *mulher_2020*, conforme mostra a Figura 4.10.

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	ficha_ano	ano	usuaria	contato	idade
1	01*2020	2020	India Viana Pimentel	5(3527)586-74-47	66
2	02*2020	2020	Sofia Boto da Costa	791(1715)084-08-41	59
3	03*2020	2020	Jane Gorjão Salomão Jan...	34(6241)194-57-42	25
4	04*2020	2020	Bianca Valadão Coimbra	60(3238)132-07-44	28
5	05*2020	2020	Ester Escobar Grande	9(717)298-74-52	32
6	06*2020	2020	Ludmila Rebouças Bugalho	187(2931)545-52-99	41
7	07*2020	2020	Ellen Leme Mirandela	83(01)200-94-51	22
8	08*2020	2020	Roberta Freiria Castelhano	7(97)803-15-86	64
9	09*2020	2020	Neusa Pacheco Matias	1(055)755-97-59	38

Figura 4.10. Tabela *usuarias_mulher_2020* consolidada, realizando a guarda das informações que devem ser mantidas em sigilo.

Finalizada a etapa de anonimização dos dados, através da separação dos dados pessoais das vítimas dos demais dados, inicia-se a etapa de limpeza dos dados.

4.4.3. Limpeza dos Dados

O processo de limpeza de dados é uma parte integrante da análise de dados como um todo e representa o processo de fixar ou remover dados formatados incorretamente, corrompidos, duplicados ou incompletos.

Durante a análise exploratória dos dados contidos na tabela *mulher_2020*, constata-se o uso de múltiplos valores para a representação de uma mesma informação dentro de uma determinada coluna da tabela, como por exemplo: a existência dos valores “ca.”, “católica”, “catolica”, “catól”, “CAT”, “catolica” e outros para representar a religião “católica” na coluna religião da tabela *mulher_2020*. Esse problema deve-se ao

fato de que o *dataset* fictício que utilizamos para este minicurso foi criado com base no *dataset* real da CEAM, em que a anotação das informações das mulheres vítimas de violência foi realizada de forma manual, por funcionários de uma instituição, ocasionando problemas de digitação e ausência de padronização dos valores.

Na experiência realizada com o *dataset* do CEAM, para resolver esse problema da padronização dos valores e garantir que a base pudesse retornar consultas SQL de modo eficiente, foi necessário realizar alguns procedimentos para a normalização dos valores das colunas da tabela. Pode-se visualizar na Figura 4.11 os diferentes valores utilizados para representar a informação da religião da usuária.

Figura 4.11. Valores distintos encontrados inicialmente na coluna "religião" do *dataset* CEAM, antes da limpeza de dados (à esquerda) e depois (à direita).

Na construção da base de dados da CEAM, o procedimento de normalização consistiu primeiro em identificar quais foram os valores distintos, usados na representação de um mesmo dado, estavam presentes em cada uma das colunas da tabela, para depois então passar à padronização desses dados. Esse procedimento foi repetido para todas as colunas da base, resultando em diminuição substancial dos valores distintos em cada uma das colunas, como se pode observar na Tabela 4.5 e na Figura 4.12.

Tabela 4.5. Redução substancial dos valores distintos para a representação dos dados existentes nas colunas da tabela *cadastro_ceam_atual* após a limpeza de dados.

	Quantidade de valores distintos encontrados nas colunas						
	origem	**bairro**	**tipos_de_vd**	**encam**	**religiao**	**grau_escolaridade**	**cor**
Antes da limpeza	140	71	150	189	43	21	21
Depois da limpeza	123	43	29	86	13	10	10
Redução	**12%**	**39%**	**81%**	**54%**	**70%**	**52%**	**52%**

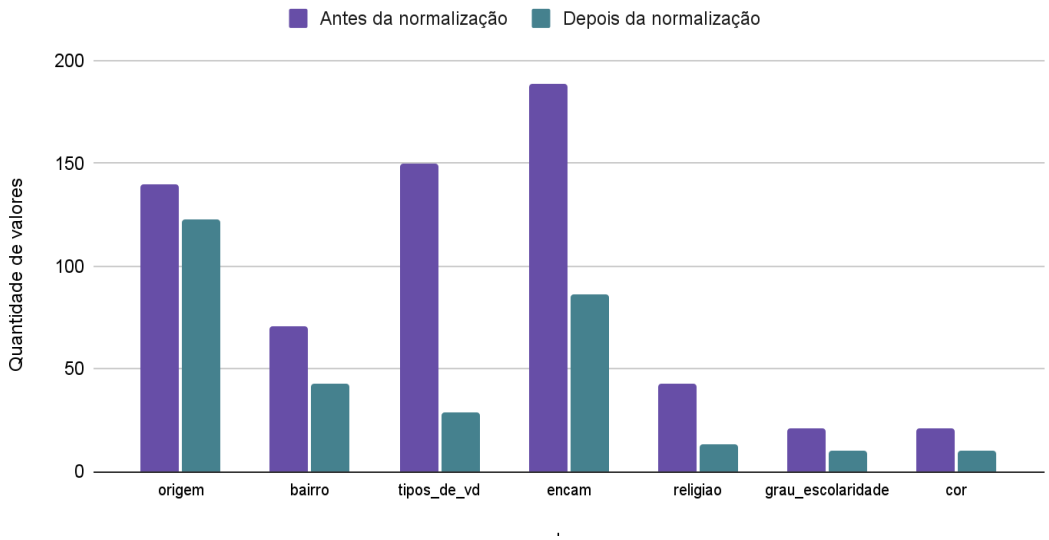

Figura 4.12. Relação entre os valores das colunas antes e depois dos procedimentos adotados para a normalização.

A transformação dos dados no contexto apresentado esteve relacionada à uma característica marcante do *dataset* real analisado, que foi a falta de padronização dos dados, em decorrência da anotação manual dos dados pela equipe técnica responsável pelos registros. A fim de tornar a tabela *mulher_2020* em um objeto próprio para a realização de consultas, procedeu-se então à padronização dos dados, através dos procedimentos descritos nas duas próximas subseções.

4.4.3.1. Identificação dos valores distintos em uma determinada coluna da tabela *mulher_2020*

A identificação dos valores distintos em uma determinada coluna, foi realizada com a instrução:

```
SELECT <coluna>, COUNT(1) AS qtd FROM mulher_2020 GROUP BY
<coluna> ORDER BY qtd DESC, <coluna>;
```

Esta consulta conta o número de ocorrências de cada valor distinto presente em uma determinada coluna da tabela. A consulta retorna esses valores distintos agrupados por ocorrência em uma coluna e em outra coluna, quantas vezes cada valor distinto apareceu. Por exemplo, ao realizar a consulta feita anteriormente na tabela *mulher_2020*:

```
SELECT religiao, COUNT(1) AS qtd FROM mulher_2020 GROUP BY
religiao ORDER BY qtd DESC, religiao;
```

O uso dessa consulta retorna uma listagem com duas colunas. Uma delas com todos os diferentes dados relativos às religiões das vítimas constantes na coluna <religiao>, ao lado da contagem da incidência de cada um dos tipos, tal como na Figura 4.13.

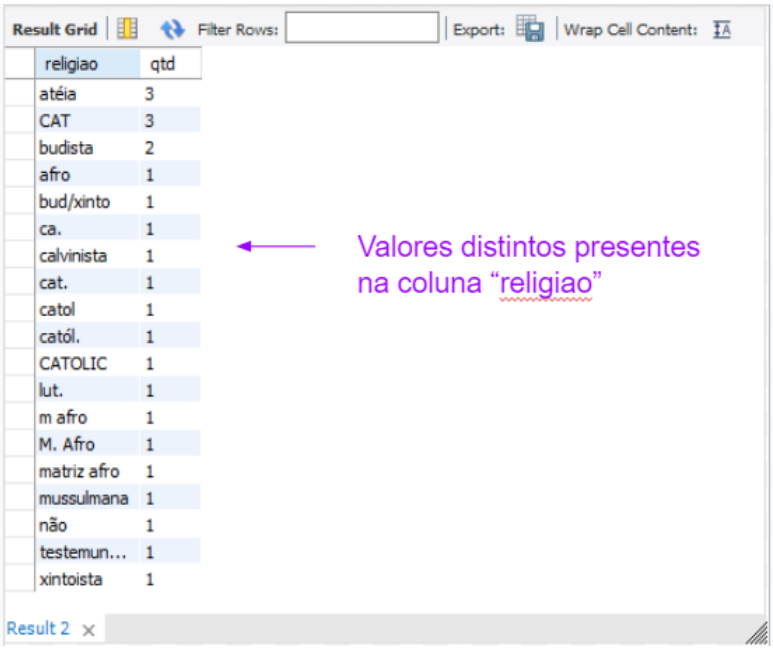

Figura 4.13. Valores distintos presentes na coluna "religião" da tabela *mulher_2020*.

4.4.3.2. Criação de um padrão correspondente ao dado que se desejava representar e padronização dos dados referentes

Esta etapa da limpeza de dados foi a que permitiu efetivamente que se pudesse trabalhar com os dados em uma fase de análise. Uma vez encontrados os valores distintos possíveis para representar um mesmo dado na tabela *mulher_2020*, criou-se um padrão para representar cada um dos valores. Por exemplo, onde se lia "CAT", "católica", "ca.", optou-se pelo padrão "catolica". Depois, passou-se para a etapa de padronização em si. Nessa etapa, utilizou-se a instrução SQL a seguir tantas vezes quantas foram necessárias para a padronização da nomenclatura de um determinado dado:

```
UPDATE mulher_2020 SET <coluna> = dado padronizado
    WHERE <coluna> = dado não padronizado;
```

Essa instrução basicamente substitui um determinado valor presente em uma coluna da tabela especificada por outro valor informado. Ao realizar as instruções UPDATE a seguir, serão padronizados os valores "CAT", "CATOLICA", "CATÓLICA" e "catól." como "catolica".

```
UPDATE      mulher_2020      SET      religiao      =      'catolica'
      WHERE religiao = 'CAT';
UPDATE      mulher_2020      SET      religiao      =      'catolica'
      WHERE religiao = 'CATOLICA';
UPDATE      mulher_2020      SET      religiao      =      'catolica'
      WHERE religiao = 'CATÓLICA';
UPDATE      mulher_2020      SET      religiao      =      'catolica'
      WHERE religiao = 'catól.';
```

Essa ação viabiliza a contagem desses valores pelas consultas SQL, uma vez que os dados estarão padronizados.

Atenção:

- Sempre que se utilizar palavras como dados da tabela durante a consulta SQL, elas deverão vir entre aspas simples (‘ ‘).
- O MySQL Workbench possui uma configuração inicial que impede as instruções que usam os comandos UPDATE e DETETE. Para configurar o sistema para aceitar essas instruções você deverá ir em <edit><preferences> e desmarcar a caixa de seleção de “*safe updates*”, como mostra a Figura 4.14.

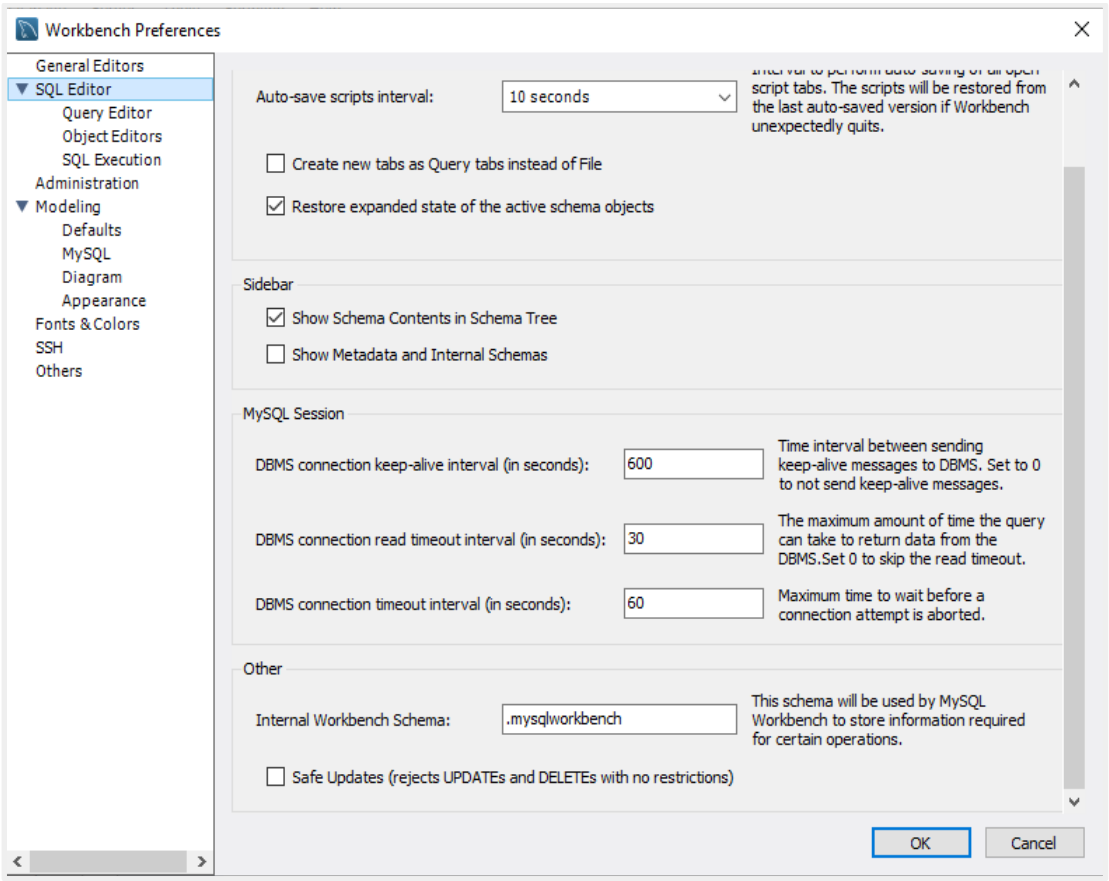

Figura 4.14. A caixa de seleção “*safe updates*” deve ser desmarcada para a realização do comando *update*.

Após o procedimento de normalização, espera-se obter uma tabela mais apropriada para a realização de análises e dados e criação de visualizações.

4.4.4. Transformando os Dados

Finalizando as etapas referentes à consolidação da base de dados para a fase de análise, é muito importante que as colunas exibam dados monovalorados. Um problema importante do *dataset mulher_2020* está na coluna <tipos_de_vd>, que relaciona os tipos de violência praticados pelos agressores, apresentando dados multivalorados, ao invés de monovalorados.

A coluna <tipos_de_vd> pode apresentar até 5 valores dentro de uma mesma célula, como por exemplo, "CORP/EMO/DIG/REC/SEX", referentes às violências "corporal", "emocional", "dignidade", "recursos/patrimônio" e "sexual". Essa disposição da informação pode dificultar muito o processamento dos dados quando carregados em um *software* de visualização de dados.

O processo de transformação dos dados representa a tarefa de modificar o formato de dados para um mais interessante ou apropriado para a mineração de dados, de maneira a agilizar o processamento de dados em bases mais extensas.

Para tornar as colunas multivaloradas em monovaloradas, foi adotada a estratégia de transformação de uma coluna multivalorada em várias colunas monovaloradas, contendo como dados 0 ou 1 (type int). No *dataset mulher_2020*, as colunas que podem receber esse tipo de transformação são a coluna <tipos_de_vd> e a coluna <encam>, que guardam os dados dos tipos de violência doméstica praticados e dos locais para onde as vítimas foram encaminhadas no momento da denúncia.

As colunas novas, com dados monovalorados, serão proporcionais aos valores existentes na coluna com dados multivalorados. Por exemplo, se a coluna <tipos_de_vd> pode apresentar os dados CORP/EMO/DIG/REC/SEX, é possível transformá-la em 5 novas colunas, as colunas <CORP>, <EMO>, <DIG>, <REC> e <SEX>, assumindo valor "0" caso o tipo de violência não esteja presente no cadastro e "1" caso o tipo de violência esteja presente no cadastro.

Para realizar a transformação, deve-se seguir os passos:

1. Identificar as colunas que possuem os dados multivalorados;
2. Identificar quais são os dados possíveis dentro de uma coluna de dados multivalorados;
3. Criar tantas novas colunas na tabela quantas forem as possibilidades de valores identificados no passo anterior;
4. Uniformização dos nomes representativos dos dados na coluna <tipos_de_vd>;
5. Representar a ocorrência (1) e a não ocorrência (0) de cada um dos dados nas novas colunas criadas.

4.4.4.1. Identificação das colunas que possuíam dados multivalorados

A identificação das colunas que possuem dados multivalorados deve se dar através da leitura livre dos dados das colunas da tabela *mulher_2020* e da percepção de dados escritos lado a lado ou mesmo separados por vírgulas, ponto e vírgula ou barras.

4.4.4.2. Identificação e anotação de todos os dados possíveis dentro de uma mesma coluna

Na coluna <tipos_de_vd>, pode se observar a olho nu, que existem 5 possibilidades de violência sofrida “CORP”, “EMO”, “DIG”, “REC” e SEX”.

4.4.4.3. Criação das novas colunas “corp”, “emo”, “dig”, “rec”, “sex” na tabela, representando as possibilidades de valores identificados na coluna multivalorada

Para criar as novas colunas aplica-se a instrução:

```
ALTER TABLE mulher_2020 ADD coluna x INT AFTER coluna y;
```

Essa instrução SQL criará uma coluna de nome <coluna x> logo após a <coluna y> informada. Para criar as colunas desejadas na tabela, basta executar as instruções que seguem. A estrutura resultante é mostrada na Figura 4.15.

```
ALTER TABLE mulher_2020
      ADD corp int AFTER rel_agressor,
      ADD emo int AFTER corp,
      ADD dig int AFTER emo,
      ADD rec int AFTER dig,
      ADD sex int AFTER rec;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

cam	dia	religiao	grau_escolaridade	cor	com_renda	filhos	rel_agressor	corp	emo	dig	rec	sex
	1/2/2020	np	ensino médio	negra	não	1	Ex marido	NULL	NULL	NULL	NULL	NULL
	1/2/2020	CAT	ensino fundamental	negra	sim	3	outros	NULL	NULL	NULL	NULL	NULL
/D/Vara	1/3/2020	luter	alfabetização	branca	sim	8	marido	NULL	NULL	NULL	NULL	NULL
_/D	1/4/2020	catolica	ensino médio	amarela	não	1	irmã	NULL	NULL	NULL	NULL	NULL
ap	1/4/2020	calvinista	ensino médio	negra	sim	2	marido	NULL	NULL	NULL	NULL	NULL
ap	1/5/2020	np	np	branca	sim	3	mãe	NULL	NULL	NULL	NULL	NULL
	1/8/2020	bud/xinto	ensino médio	amarela	sim	3	marido	NULL	NULL	NULL	NULL	NULL
/Médico de família	1/8/2020	budista	alfabetização	amarela	sim	0	desconhecido	NULL	NULL	NULL	NULL	NULL
ap	1/10/2020	atéia	ensino fundamental	mestiça	sim	2	nora	NULL	NULL	NULL	NULL	NULL
ap	1/12/2020	CATOLICA	alfabetização	amarela	sim	1	np	NULL	NULL	NULL	NULL	NULL
	1/15/2020	np	alfabetização	negra	np	7	marido	NULL	NULL	NULL	NULL	NULL
ap	1/16/2020	CA	alfabetização	mestiça	sim	2	marido	NULL	NULL	NULL	NULL	NULL
	1/22/2020	M. Afro	ensino médio	negra	sim	4	companheiro	NULL	NULL	NULL	NULL	NULL
istência social	1/23/2020	ca.	ensino médio	branca	não	2	companheiro	NULL	NULL	NULL	NULL	NULL
	1/23/2020	CATÓLICA	np	branca	não	3	tio	NULL	NULL	NULL	NULL	NULL

Figura 4.15. Colunas “corp”, “emo”, “dig”, “rec” e “sex”, criadas após a coluna rel_agressor.

4.4.4.4. Colocando todos os dados da coluna tipos_de_vd em letra maiúscula

Para tratar na padronização da representação dos dados na coluna, a fim de facilitar as próximas transformações que serão feitas na tabela, é indicado colocar todos os dados em caixa alta ou em caixa baixa. Para colocá-los em caixa alta, realiza-se a seguinte instrução:

```
UPDATE mulher_2020 SET tipos_de_vd = UPPER(tipos_de_vd);
```

4.4.4.5. Identificando os diferentes possíveis valores para a coluna tipos_de_vd e instanciação das colunas geradas

O próximo passo é descobrir quais são os valores distintos possíveis na coluna <tipos_de_vd>. Isto porque, a partir dessa descoberta, será possível compreender quais novas instruções serão necessárias para levar a informação da coluna multivalorada <tipos_de_vd> para as novas colunas criadas. Para saber quais são os valores distintos da coluna <tipos_de_vd>, realiza-se a seguinte consulta:

```
SELECT DISTINCT tipos_de_vd FROM mulher_2020 order by tipos_de_vd;
```

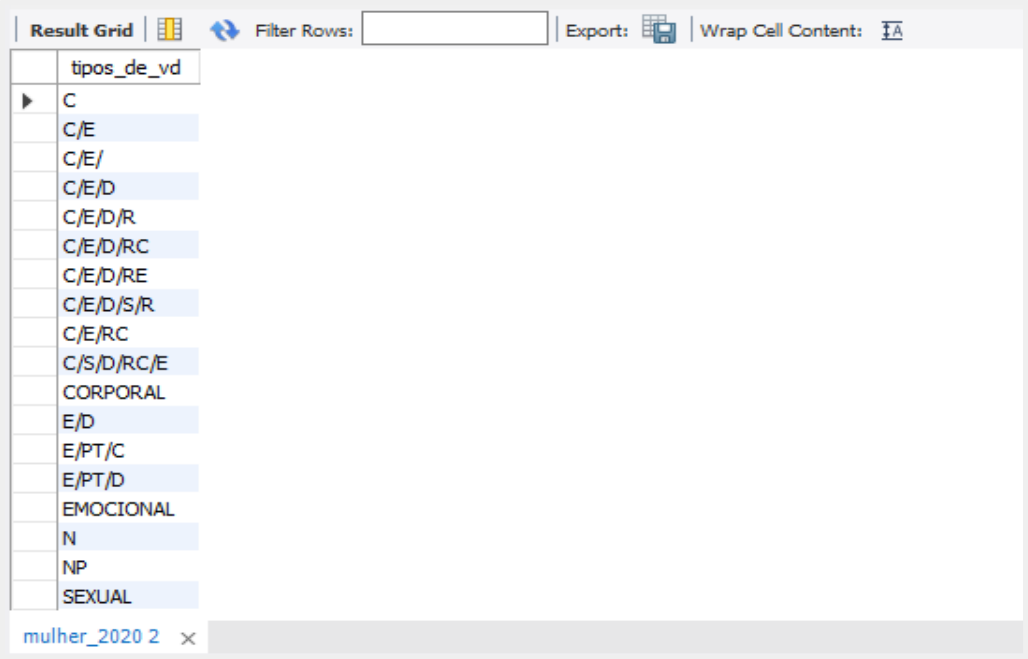

Figura 4.16. Distintas possibilidades de valores na coluna <tipos_de_vd>.

A partir da identificação dos valores possíveis na coluna <tipos_de_vd>, a ação seguinte se concentra em preencher as colunas geradas com 1, iniciando pela coluna <corp>. Observando a Figura 4.16, percebe-se que todas as células que possuem a informação de que houve violência corporal, possuem a letra “C” como indicação. Entretanto, como a letra “C” está vindo seguida de diversas informações dentro das células, será necessário distinguir que só queremos colocar 1 na coluna <corp> para as células cuja informação se inicia com a letra “C”, podendo ser seguida de qualquer outro

dado. Para realizar essa representação, utiliza-se o símbolo de porcentagem (%) logo após a letra “C”. Assim: ‘C%’, não esquecendo das aspas simples para indicar que se trata de uma *string*.

```
UPDATE mulher_2020 SET corp = 1 WHERE tipos_de_vd LIKE 'C%';
```

Semelhante modo deve ser feito para preencher a coluna <corp> com 1 quando o dado da coluna <tipo_de_vd> for igual a *string* ‘CORPORAL’.

```
UPDATE mulher_2020 SET corp = 1 WHERE tipos_de_vd LIKE 'CORPORAL';
```

Procedimentos semelhantes serão aplicados para a construção das outras colunas. Entretanto, é importante observar que a análise para a definição da melhor instrução deverá ser feita caso a caso.

Para preencher a coluna <emo> com 1 quando o dado na coluna <tipos_de_vd> começar com a letra “E”, contiver a letra “E” entre duas barras (/), contiver a string “EMOCIONAL” ou terminar com uma barra (/) seguida da letra “E” usa-se as instruções:

```
UPDATE mulher_2020 SET emo = 1 WHERE tipos_de_vd LIKE 'E%';
UPDATE mulher_2020 SET emo = 1 WHERE tipos_de_vd LIKE '%/E/%';
UPDATE mulher_2020 SET emo = 1 WHERE tipos_de_vd LIKE 'EMOCIONAL';
UPDATE mulher_2020 SET emo = 1 WHERE tipos_de_vd LIKE '%/E';
```

Segue a mesma lógica para preencher a coluna <dig> com 1, quando o dado na coluna <tipos_de_vd> terminar com “/D” ou possuir a letra “D” entre duas barras (/D/).

```
UPDATE mulher_2020 SET dig = 1 WHERE tipos_de_vd LIKE '%/D';
UPDATE mulher_2020 SET dig = 1 WHERE tipos_de_vd LIKE '%/D/%';
```

As instruções que seguem efetivam o preenchimento da coluna <rec> com 1, quando o dado na coluna <tipos_de_vd> terminar com “/R”, terminar com “/RC”, terminar com “/RE”, contiver “/RC/” e quanto contiver “/PT/”.

```
UPDATE mulher_2020 SET rec = 1 WHERE tipos_de_vd LIKE '%/R';
UPDATE mulher_2020 SET rec = 1 WHERE tipos_de_vd LIKE '%/RC';
UPDATE mulher_2020 SET rec = 1 WHERE tipos_de_vd LIKE '%/RE';
UPDATE mulher_2020 SET rec = 1 WHERE tipos_de_vd LIKE '%/RC/%';
UPDATE mulher_2020 SET rec = 1 WHERE tipos_de_vd LIKE '%/PT/%';
```

Para preencher a coluna <sex> com 1, quando o dado na coluna <tipos_de_vd> for igual à string “SEXUAL” use a próxima instrução:

```
UPDATE mulher_2020 SET sex = 1 WHERE tipos_de_vd LIKE 'SEXUAL';
```

Depois que terminar essa etapa de preenchimento das colunas, elas terão os valores “1” e “null” como dados. Então será o momento de trocar os valores “null” por “0”, para obter o formato desejado no planejamento da transformação de dados. As próximas instruções de UPDATE realizam a alteração desejada.

```
UPDATE mulher_2020 SET corp = 0 WHERE corp is null;
UPDATE mulher_2020 SET emo = 0 WHERE emo is null;
UPDATE mulher_2020 SET dig = 0 WHERE dig is null;
UPDATE mulher_2020 SET rec = 0 WHERE rec is null;
UPDATE mulher_2020 SET sex = 0 WHERE sex is null;
```

O resultado dessas consultas podem ser visualizados na Figura 4.17.

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

le_vd	encam	dia	religiao	grau_escolaridade	cor	com_renda	filhos	rel_agressor	corp	emo	dig	rec	sex
	np	1/2/2020	np	ensino médio	negra	não	1	Ex marido	0	0	0	0	0
	T/C	1/2/2020	CAT	ensino fundamental	negra	sim	3	outros	1	1	1	1	0
	T/A/D/Vara	1/3/2020	luter	alfabetização	branca	sim	8	marido	1	1	1	0	0
.E	DEL/D	1/4/2020	catolica	ensino médio	amarela	não	1	irmã	1	1	1	1	0
.C	terap	1/4/2020	calvinista	ensino médio	negra	sim	2	marido	1	1	1	1	0
	terap	1/5/2020	np	np	branca	sim	3	mãe	1	1	1	1	0
	np	1/8/2020	bud/xinto	ensino médio	amarela	sim	3	marido	1	1	1	1	0
.E	AM/Médico de família	1/8/2020	budista	alfabetização	amarela	sim	0	desconhecido	1	1	1	1	0
	Terap	1/10/2020	atéia	ensino fundamental	mestiça	sim	2	nora	0	0	0	0	0
	Terap	1/12/2020	CATOLICA	alfabetização	amarela	sim	1	np	1	0	0	0	0
	A/T	1/15/2020	np	alfabetização	negra	np	7	marido	1	1	0	1	0
	terap	1/16/2020	CA	alfabetização	mestiça	sim	2	marido	1	1	1	0	0
.C/E	ter	1/22/2020	M. Afro	ensino médio	negra	sim	4	companheiro	1	1	1	1	0
)NAL	Assistência social	1/23/2020	ca.	ensino médio	branca	não	2	companheiro	0	1	0	0	0
	np	1/23/2020	CATÓLICA	np	branca	não	3	tio	0	0	0	0	0

Figura 4.17. Formato das colunas “corp”, “emo”, “dig”, “rec” e “sex” depois do processo de transformação.

4.4.5. Resultados da Base Consolidada

Como resultado do processo de aquisição e transformação dos dados, obteve-se a base de dados nomeada *violencia_mulher*, com as tabelas *mulher_2020* e *usuarias_mulher_2020*, relacionadas através da coluna <id>. A base consolidada permite a etapa da integração de dados, que é o carregamento de dados, que tem como objetivo gerar valor sobre eles para a formação de relatórios e análises capazes de dar suporte à tomada de decisão. Um exemplo de etapa de integração será mostrado na seção a seguir, com a utilização da base consolidada para a criação de um *dashboard* interativo.

4.5. Sumarização de Dados com *Dashboards*

Nesta seção, serão descritas algumas técnicas e práticas relacionadas à sumarização de dados através de *dashboards*, passando pela criação de perguntas de pesquisa a serem feitas diante da base de dados consolidada até as técnicas geração de *dashboards* e as decisões de *design* que permeiam todo o processo.

4.5.1. Como Visualizar os Dados?

A partir da base gerada dos dados, a análise dos dados buscou responder questionamentos que as especialistas do CEAM precisam submeter em relatórios, a fim de simular perguntas de pesquisa que dariam suporte para gestores/tomadores de decisão no âmbito das políticas públicas para população em estado de vulnerabilidade social — Mulheres. Desta forma, foram levantadas as perguntas:

- Qual o número de registros coletados?
- Qual é a idade média das vítimas?
- Qual é o número médio de filhos das vítimas?
- Qual é o perfil da distribuição da quantidade de filhos das vítimas de agressão?
- Qual é a proporção entre as raças das usuárias?
- Quais são os graus de parentesco entre as vítimas e os agressores?
- Quais são os tipos de agressões mais praticadas? (não deixando de considerar que a base de dados informa que as vítimas informam na grande maioria das vezes sofrer mais de um tipo de violência, de forma cumulativa).
- Qual é a distribuição da quantidade de vítimas agredidas por distritos do município?
- Qual é a incidência de agressões considerando a divisão entre mulheres que possuem renda e as que não possuem renda?

Na próxima seção, serão realizadas análises a fim de encontrar respostas para as perguntas listadas. Através da elaboração de um *dashboard* para visualização dos dados, serão apresentadas as ferramentas e metodologias que podem contribuir para a realização da sumarização de dados e informação dos resultados das análises para a audiência.

4.5.2. Decisões Técnicas

A forma de sumarização dos dados escolhida para criar uma visualização com os dados da CEAM foi a criação de um *dashboard*, de maneira que as perguntas de pesquisa levantadas na seção 4.4.1 puderam ser filtradas e cruzadas. Esse *dashboard* foi concebido através da ferramenta Google Data Studio, que possibilita um ambiente colaborativo diretamente integrado com a plataforma Google Sheets, para onde os dados poderão ser posteriormente exportados.

Os dados disponibilizados no *dashboard* foram separados em duas páginas, a primeira contendo as informações principais discutidas na seção 4.4.1, e a segunda

possuindo um enfoque maior nos tipos de violência ocorridos. Para facilitar a análise através de alguns critérios, foram implementados filtros como o ano da agressão e a raça da mulher agredida.

4.5.3. Gerando o *Dashboard*

Para gerar um ambiente que proporcione a visualização dos dados, é necessário gerar um novo relatório na ferramenta Google Data Studio. Após fazer o login com uma conta Google, todos os seus relatórios podem ser acessados através da página Google Data Studio[12].

Uma das grandes vantagens da ferramenta Google Data Studio é a integração facilitada com os dados presentes na ferramenta de armazenamento da nuvem Google Drive. Selecionando a opção "*Blank Report"*, é possível importar a base de dados gerada, que deverá ser armazenada em uma planilha na ferramenta Google Sheets, conforme a Figura 4.18.

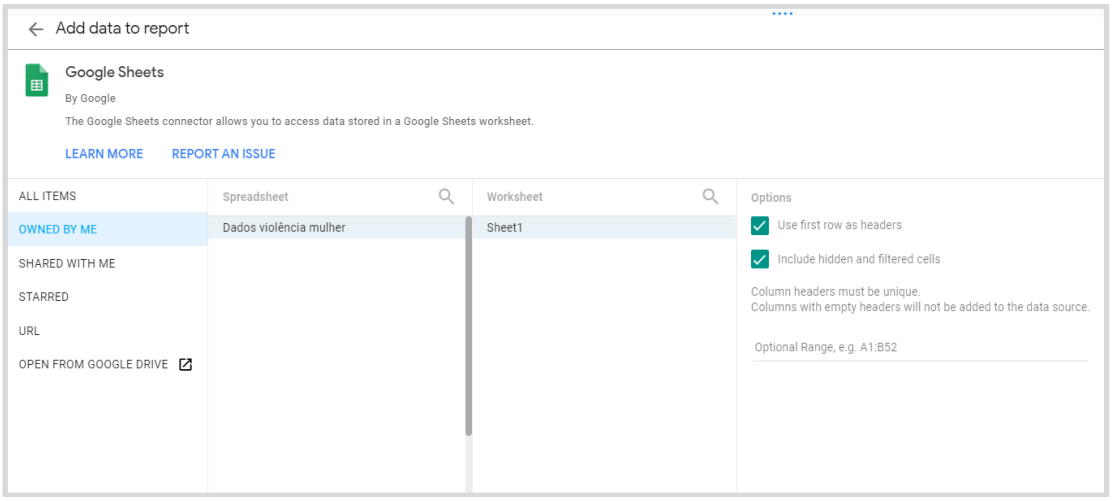

Figura 4.18. Importação da base de dados na ferramenta Google Data Studio.

Após importar a base de dados, será possível inserir a primeira visualização. Para saber a quantidade total de registros coletados, será gerada uma visualização de *scorecard* para mostrar estes dados. Essas opções de visualização podem ser vistas a partir da opção "*Add a Chart*", localizada no menu superior da ferramenta, conforme a Figura 4.19.

[12] https://datastudio.google.com/. Acesso em 29 de agosto de 2021.

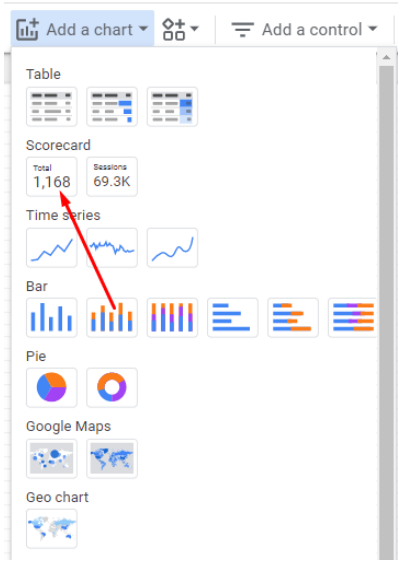

Figura 4.19. Opções de visualização.

Ao gerar uma visualização, um menu à direita será disponibilizado, mostrando alguns dos atributos principais que podem ser alterados para que sejam fornecidas as informações desejadas. Estes atributos estão presentes na Figura 4.20, seguida de uma breve descrição das suas características individuais principais.

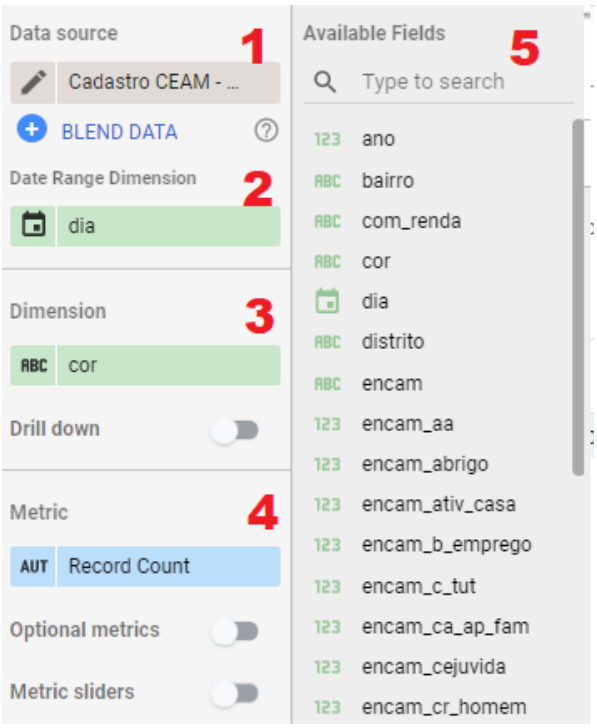

Figura 4.20. Os atributos principais em uma visualização de dados no Google Data Studio: *Data Source, Date Range Dimension, Dimension, Metric e Available Fields.*

4.5.3.1. *Data Source*

Representa a fonte de dados, ou seja, a base a partir da qual será obtida essa métrica. Caso a análise seja feita olhando em duas bases ou mais, é importante sempre verificar qual é a base a ser consultada.

4.5.3.2. *Date Range Dimension*

A dimensão de intervalo temporal informa ao Data Studio quais das dimensões disponíveis controlam o intervalo de tempo a ser trabalhado. Normalmente, este campo é preenchido automaticamente por campos que possuem a tipagem de data, por conseguirem auxiliar a determinar um intervalo para a análise.

4.5.3.3. *Dimension*

Este campo determina a dimensão na qual as métricas serão baseadas. Uma dimensão serve para descrever e categorizar os seus dados. Dessa forma, os dados serão agrupados dado algum campo único que será utilizado, garantindo a individualidade dos registros.

4.5.3.4. *Metric*

Este campo determina e mensura as métricas que serão analisadas na visualização. Uma "métrica" é o resultado de uma aplicação ou agregação de um conjunto de valores. Dentro deste campo, é possível fazer análises de seus valores através de algumas medidas matemáticas, como somatório, média ou número total de registros.

Neste campo, também é possível editar o nome do campo a ser mostrado na métrica, clicando no ícone de editar, logo acima da métrica. Isto será especialmente útil para inserir nomes mais intuitivos na visualização, conforme a Figura 4.21.

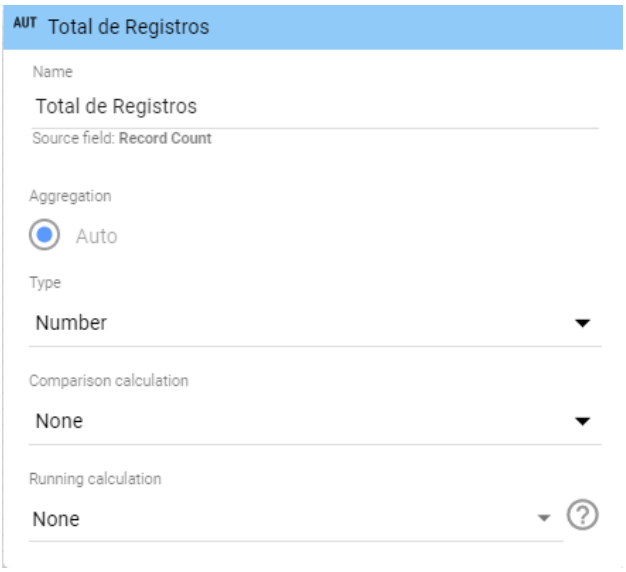

Figura 4.21. Edição do nome de métrica.

4.5.3.5. *Available Fields*

Nesta área, todas as colunas da base de dados importada são convertidas em campos, que podem ser modelados como dimensões ou métricas a serem analisadas. Estes campos serão a base para gerar funções, operadores e inserir a lógica nos dados. Cada campo possui uma tipagem diferente, podendo ser caracterizadas como strings, booleanos ou valores numéricos, por exemplo.

4.5.4. Configurando e Gerando Novas Visualizações

Ao configurar o *Scorecard*, é possível observar que a dimensão é categorizada como o campo "Dia", pois a data possui dados únicos para agregar os registros. A métrica em si será o campo numérico "*Record Count*", gerado automaticamente pelo Google Data Studio para contabilizar o valor geral de todos os registros presentes na importação. Por fim, será necessário editar o nome da métrica, substituindo para "Total de Registros".

Partindo para outra visualização, desta vez será mostrada a "proporção racial das mulheres". Para ilustrar melhor esses dados, será selecionada a opção do Gráfico de Pizza, localizado conforme a Figura 4.22.

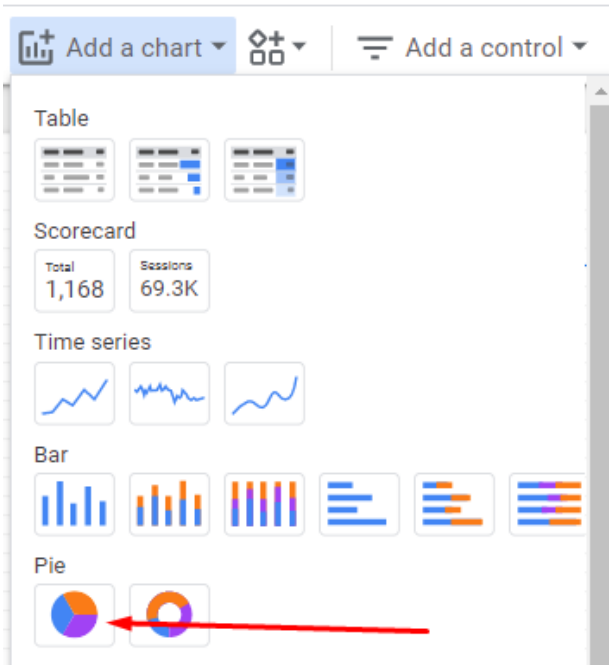

Figura 4.22. Opção da visualização de gráficos de setores ou de pizza (*Pie chart*).

Para configurar o gráfico, é preciso mudar a dimensão padrão que está sendo utilizada. Assim, a dimensão atual deverá ser alterada para o campo "cor", que representa as raças reportadas. A métrica será o total de registros, pois será necessária a proporção de respostas de acordo com a quantidade de itens especificados na dimensão. O resultado final pode ser visto na Figura 4.23.

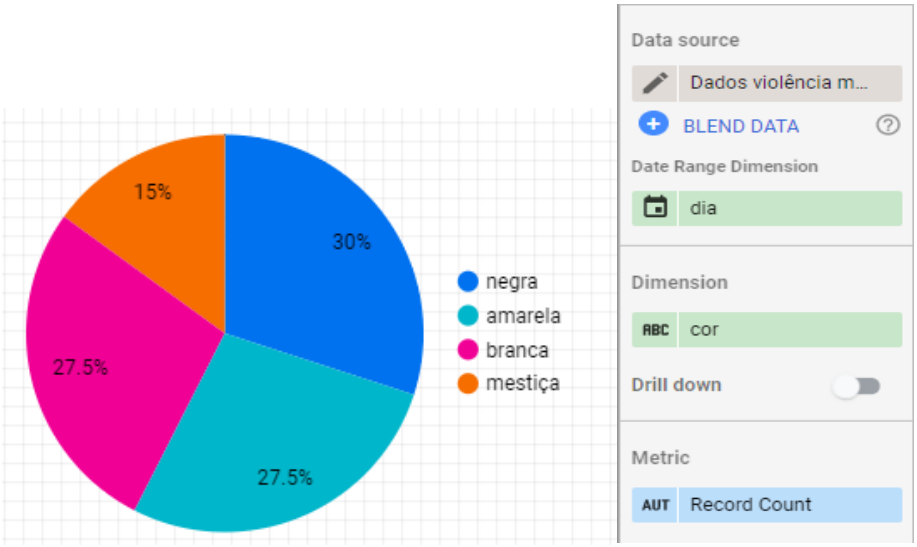

Figura 4.23. Gráfico de pizza gerado, seguido de suas propriedades.

Para finalizar as visualizações, será gerado um gráfico para mostrar a relação do agressor com a vítima. Para ilustrar melhor esses dados, será selecionada a opção do Gráfico de barras em colunas, localizado conforme a Figura 4.24.

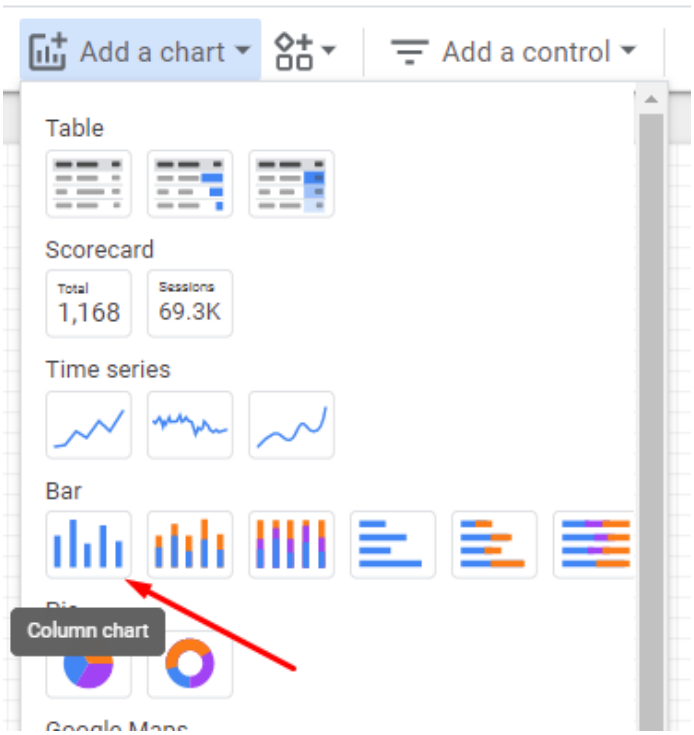

Figura 4.24. Opção da visualização de gráficos de barras em colunas.

Para configurar o gráfico, é preciso novamente mudar a dimensão padrão que está sendo utilizada. Assim, a dimensão atual é alterada para o campo “rel_agressor”, que representa o grau de parentesco do agressor. A métrica continuará sendo o total de registros. O resultado final pode ser visto na Figura 4.25.

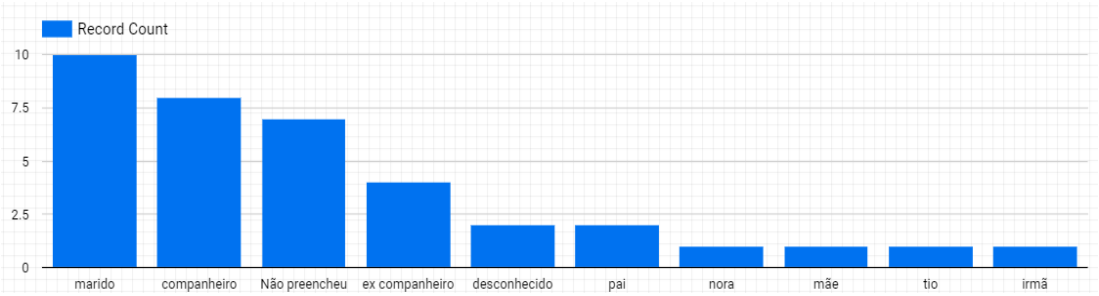

Figura 4.25. Gráfico de colunas gerado, seguido de suas propriedades.

Caso haja a necessidade de mudar a cor do gráfico de barras, é possível mudar dentro da aba lateral de propriedades da visualização. Na aba *Style*, será mudada a cor das barras de azul para vermelho, conforme Figura 4.26.

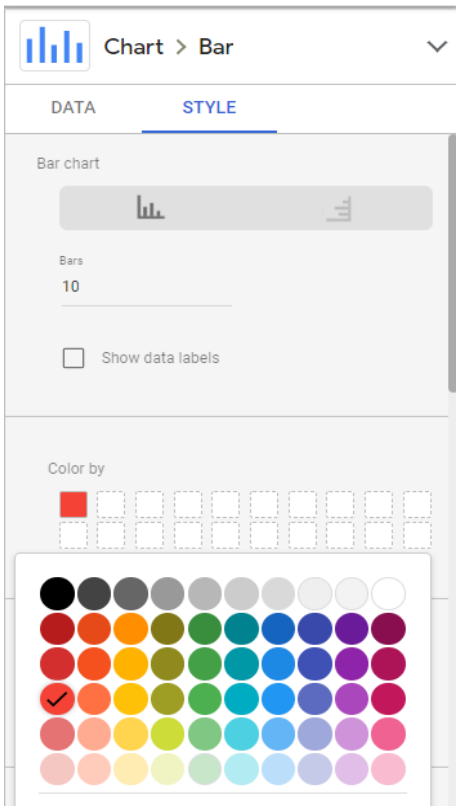

Figura 4.26. Gráfico de colunas gerado, seguido de suas propriedades.

A página é então finalizada com três visualizações principais: A quantidade de registros, a percentagem de raças informadas e a comparação entre os agressores reportados.

4.5.5. Decisões de Design

Dado que a visualização gerada trabalha com dados sensíveis, é extremamente importante garantir que as informações sejam transmitidas de maneira que se consiga capturar a atenção e despertar os sentimentos corretos das pessoas que irão interagir com ela. Dessa forma, a implementação da identidade visual foi concebida através de um conjunto de regras de design durante a montagem do *dashboard*.

4.5.5.1. Escolha das cores

Uma das técnicas implementadas se dá pelo uso correto das cores a serem implementadas no *dashboard*. Goethe (1967) iniciou os primeiros estudos voltados ao efeito da influência das cores na percepção do ser humano. A área atual, denominada de psicologia das cores, aborda um conjunto de sentimentos e sensações que determinada cor pode transparecer, de acordo com o contexto em que ela está sendo abordada.

Considerando o contexto de violência e urgência explicitados nos dados, foram escolhidas três cores principais para compor o *dashboard*: a cor preta, compondo o sentimento de pesar e luto, a cor vermelha, compondo o sentimento de perigo e urgência e a cor branca, simbolizando traços de simplicidade e honestidade dos dados. A paleta de cores escolhida está dentro dos conformes de acessibilidade de acordo com Web Content Accessibility Guidelines (WCAG) [Caldwell, Cooper, Reid, Gregg *et al*. 2008], garantindo também a facilidade na leitura dos dados através de dispositivos de diferentes resoluções.

4.5.5.2. Proporção entre as cores

Com a paleta de cores escolhida, também foi necessário definir onde as mesmas seriam aplicadas. Uma das técnicas principais utilizadas para facilitar essa tomada de decisão foi o princípio 60-30-10, que divide uma paleta de três cores a serem utilizadas de acordo com essa proporção no projeto. A cor primária, compondo aproximadamente 60% da composição, foi a cor preta, representando o fundo utilizado na interface. Já alguns detalhes e gráficos foram montados com a cor secundária vermelha, representando 30% da composição na interface. Por último, foi utilizada a cor branca para destacar alguns detalhes e legendas nos gráficos montados, representando 10% da composição.

4.5.6. Estilizando a Página

Seguindo o princípio do 60-30-10, será criado um fundo preto para o nosso *dashboard*. Sem selecionar nenhum gráfico, o menu da lateral direita do Data Studio irá mostrar algumas opções de Temas e *Layout* customizados. Na aba “Theme”, clique no botão *Customize* e troque a cor do fundo em “*Report Background*” para preto. O resultado pode ser visto conforme a Figura 4.27.

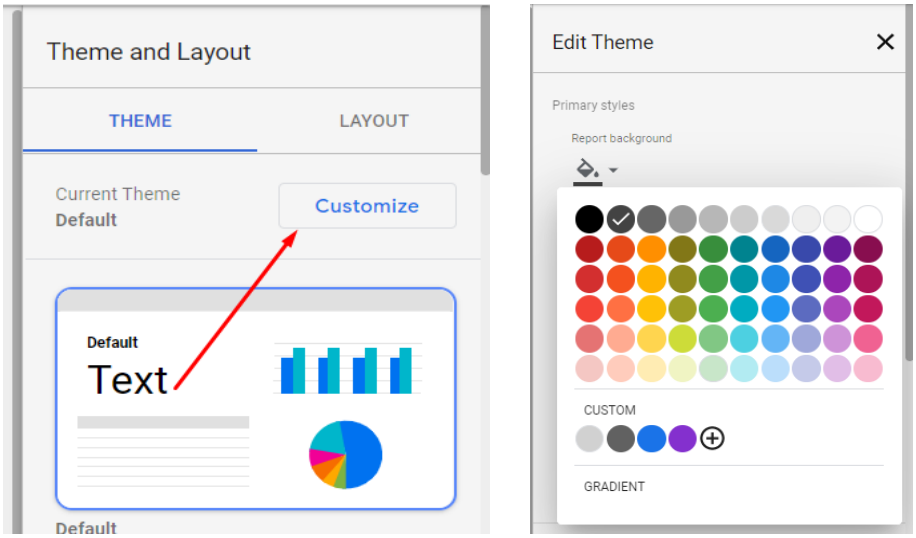

Figura 4.27. Customização de cor base do *dashboard*.

É possível também alterar a fonte e a cor geral dos textos. Na mesma aba onde foi alterado o fundo, mude a fonte para “Oswald” e a cor para branco, conforme a Figura 4.28.

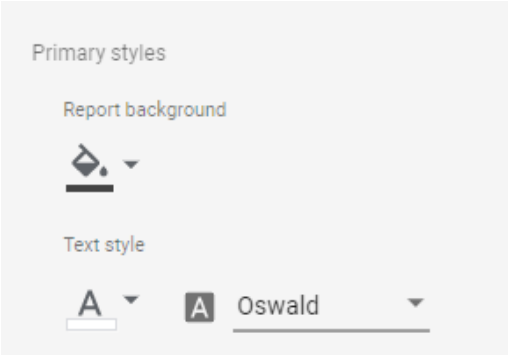

Figura 4.28. Customização de texto base do *dashboard*.

O próximo passo é gerar algumas divisórias e textos para os gráficos em si. Criando formas geométricas, é possível usar isso como uma espécie de container para as visualizações.

Crie um retângulo clicando no ícone no *menu* principal, e insira as devidas propriedades para dar fundo, borda e cor à forma, conforme a Figura 4.29.

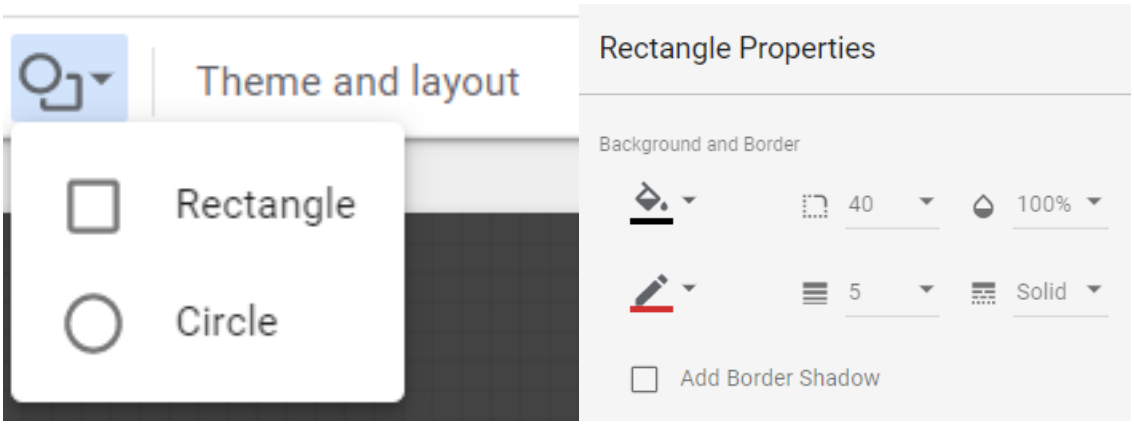

Figura 4.29. Criando formas e alterando as suas propriedades de borda e cor.

É possível então inserir uma caixa de texto dentro das nossas divisórias, de modo a ilustrar melhor as visualizações em si. Um título também será inserido do lado de fora, para representar o nome da pesquisa. Clique no campo de inserir Texto e insira os textos conforme a Figura 4.30 e ajuste-os para chegar em algo similar à Figura 4.31.

Figura 4.30. Criando textos.

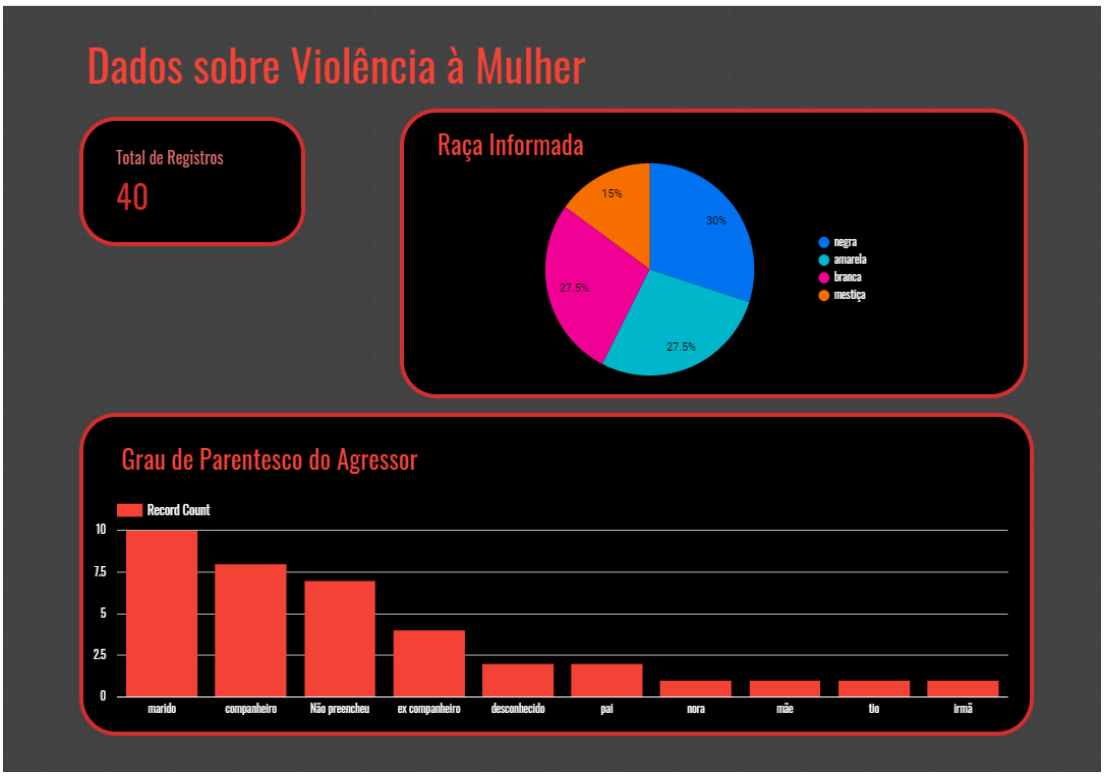

Figura 4.31. Produto final: a visualização gerada.

4.6. Considerações Finais

Neste capítulo foram abordadas técnicas e práticas de Jornalismo de Dados tais como coleta de dados, anonimização de bases de dados, limpeza e transformação de dados utilizando um banco de dados MySQL como ferramenta e sumarização de informações por meio de um *dashboard* interativo. Para isto foi utilizado como evento propulsor, um caso real de jornalismo de dados no contexto da violência contra a mulher, que se manifestou durante uma pesquisa realizada pelo grupo de literacia de dados do PPGI da UFRJ com dados sigilosos do Centro Especializado de Atendimento à Mulher de um município do Estado do Rio de Janeiro.

Espera-se que este material seja útil não somente para divulgar conhecimento sobre os recursos e ferramentas disponíveis para a atividade jornalística por meio de dados, como também para melhorar competências e atitudes relacionadas à literacia de dados de estudantes e pesquisadores brasileiros, sensibilizando-os para a temática da violência contra a mulher e abrindo portas para uma relação ainda mais integrada dos cidadãos com uma das demandas sociais mais urgentes da humanidade, que é o combate à violência contra a mulher.

Referências

ABRAJI (2012) "Jornalismo Investigativo - definições de associados e seguidores". Disponível em: https://abraji.org.br/noticias/jornalismo-investigativo-definicoes-de-associados-e-seguidores. Acesso em agosto de 2021.

Alves, M. C. L., Dumaresq, M. L. and Silva, R. V. (2016) "As Lacunas no Enfrentamento à Violência contra a Mulher: análise dos bancos de dados existentes acerca da vigilância doméstica e familiar". Brasília: Núcleo de Estudos e Pesquisas Senado. Disponível em https://www2.senado.leg.br/bdsf/handle/id/519161. Acesso em março de 2021.

Appelgren, E. and Nygren G. (2014) "Data journalism in Sweden: Introducing new methods and genres of journalism into 'old' organizations", Digital journalism, v. 2, n. 3, p. 394-405.

Bazzi, C. L. (2013) "Introdução a Banco de Dados". Editora UFTPR, Curitiba.

Bioni, B. R. (2019) "Proteção de dados pessoais: a função e os limites do consentimento" Rio de Janeiro: Forense.

Borburema, T. L. R., Pacheco, A. P., Nunes, A. A., Moré, C. L. O. O. and Krenkel S. (2017). "Violência contra mulher em contexto de vulnerabilidade social na Atenção Primária: registro de violência em prontuários". Revista Brasileira de Medicina de Família e Comunidade.;12(39):1-13. http://dx.doi.org/10.5712/rbmfc12(39)1460.

Bounegru, L., Chambers, L. and Gray, J. (2012) "The Data Journalism Handbook 1", Produced by European Journalism Centre, Disponível em: https://datajournalism.com/read/handbook/one. Acesso em Agosto de 2021.

Brasil. (2006) "Lei Maria da Penha". Lei nº. 11.340, de 7 de agosto de 2006.

Brasil. (2018) "Lei Geral de Proteção de Dados Pessoais". Lei nº13.709, de 14 de agosto de 2018.

Burelomova, A. S., Gulina, M. A. and Tikhomandritskaya, O. A. (2018) “Intimate partner violence: An overview of the existing theories, conceptual frameworks, and definitions”, Psychology in Russia: State of the art, v. 11, n. 3, p. 128-144.

Caldwell, B., Cooper, M., Reid, L. G., Gregg Vanderheiden *et al*. (2008) Web Content Accessibility Guidelines (WCAG) 2.0. W3C Recommendation 11 December 2008. Disponível em https://www.w3.org/TR/WCAG20/. Acesso em Maio de 2021.

Ferreira F. (2012) “Afinal, o que é o jornalismo?” Disponível em: http://www.observatoriodaimprensa.com.br/feitos-desfeitas/ed719-afinal-o-que-e-jornalismo/. Acesso em agosto 2021.

Goethe W. (1810/1967). “Theory of Colours”. London: Frank Cass.

GovLab (2021) “Data Collaboratives”. Disponível em https://datacollaboratives.org/. Acesso em agosto de 2021.

Michaelis (2021) “Dicionário Brasileiro da Língua Portuguesa” Disponível em https://michaelis.uol.com.br/busca?r=0&f=0&t=0&palavra=an%C3%B4nimo. Acesso em abril de 2021.

Open Knowledge Foundation (2021) “Open Definition”. Disponível em http://opendefinition.org/. Acesso em agosto de 2021.

Oracle (2021) “Banco de dados definido” Disponível em: https://www.oracle.com/br/database/what-is-database/. Acesso em agosto de 2021.

Oracle Corporation and/or Its Affiliates (2021) “Why MySQL?” Disponível em: https://www.mysql.com/why-mysql/. Acesso em agosto de 2021.

Villars, R. L., Olofson, C. W. and Eastwood, M. (2011) “Big data: What it is and why you should care”, White paper, IDC, v. 14, p. 1-14.

World Health Organization *et al*. (2020) “COVID-19 and violence against women: what the health sector/system can do”, 7 April 2020. World Health Organization.

Luciana Sá Brito é Doutoranda em Informática/Gestão de Sistemas Complexos (UFRJ), Mestre em Informática (UFRJ/2020), Especialista em EAD/Educação Profissional (SENAC/2012), Licenciada em Física (UFRJ/2008) e Técnica em Artes Dramáticas (ETE Martins Pena/2010). Luciana atua como Designer Instrucional na Fundação CECIERJ desde 2014 e foi Física Docente na SEEDUC-RJ (2007-2014). Seus interesses principais estão relacionados à pesquisa em Design Instrucional e Sistemas Colaborativos, pesquisa em Interação Humano Computador e à Ciência de Dados como forma de ativismo. Nas horas livres pratica montanhismo e toca piano.

Alayne Duarte Amorim é Doutoranda em Computação pela UFRJ e Mestre em Informática pelo UFRJ/2012. Possui especialização em Gerência de Tecnologia de Computação pela UFF/2003 e especialização em Administração de Banco de Dados pela Universidade Estácio de Sá em 2006. Licenciada para disciplinas de informática para o ensino fundamental e médio pelo Instituto a Vez do Mestre (2009) e graduada em SI pela Universidade do Grande Rio (1999). É professora e coordenadora pedagógica do curso técnico integrado em Informática no Colégio Pedro II.

André Viana Tardelli é Mestrando em Informática pelo grupo GRECO/PPGI da Universidade Federal do Rio de Janeiro e coordenador dos cursos de User Experience (UX) E Design no Grupo Alura. Também atua como instrutor na Caelum, ministrando aulas dos cursos de UX, Desenvolvimento Web e de Python voltado para Ciência da Dados.

Angélica Fonseca da Silva Dias é Doutora em Informática pelo PPGI/UFRJ. Atualmente é Diretora do Instituto Tércio Pacitti-NCE/UFRJ, Docente colaborador da UFRJ e do SEMESP/MEC. Coordenadora do Lab. De Pesq. Computacionais em Economia Circular voltados para os ODS-Agenda 2030. Membro Consultivo da Comissão Especial de Sistemas Colaborativos (CE-SC) – SBC. Tem experiência na área de Sistemas de Informação. Mãe de duas filhas lindas. Ama ler, pilates e uma boa corrida.

Juliana Baptista dos Santos França é Doutora em Informática pelo PPGI/UFRJ (2018) e concluiu seu Pós-doc no PPGI/UFRJ em 2019. Atualmente é docente no IC/UFRJ. Atua no programa de pós-graduação de Gestão e Estratégia (PPGE/UFRRJ) como membro permanente desde 2021 e colabora com o PPGI/UFRJ desde 2018. É membro da Comissão Especial de Sistemas Colaborativos (CE-SC) da SBC desde 2019 e também da ACM/FCA no mandato de 2019-2021. Seu principal interesse está na área de CSCW, com contribuições aos domínios de Suporte à Decisão, Gestão por Processos de Negócio, Crises e Desastres, e Informática na Educação, mas não limitado a estes.

Adriana Santarosa Vivacqua é Professora do Instituto de Computação da UFRJ. Obteve o doutorado na COPPE/UFRJ em (2007). Foi bolsista de Produtividade do CNPq e Jovem Cientista da FAPERJ. Seus interesses de pesquisa incluem IHC inteligente, CSCW, visualização. Atua como Chair Associado para Equidade junto ao Comitê Executivo da ACM SIGCHI, e foi eleita recentemente para a posição de VP-at-Large no Comitê Executivo da ACM SIGCHI (mandato 2021-2024).