Trabalho Final - Parte 2
Relatório Final
Data Warehousing e Inteligência de Negócio - 2024/2

Daniel Arruda Ponte[1], **Igor Santos**[1], **Matheus da Cruz Percine Pinto**[1]

[1] Instituto de Computação – Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil.

{danielap, igorjs, matheuscpp}@dcc.ufrj.br

1. Introdução

O tema do nosso trabalho é o perfil demográfico dos docentes nas IES e o comparativo com base na população e área dos municípios brasileiro das IES. Considerando nosso escopo, foram removidas variáveis que consideramos irrelevantes para análise; entre elas podemos citar os códigos, as mantenedoras, as bibliotecas, as mesorregiões das IES já que eles não acrescentam ou pouco acrescentam de informação útil para nossas análises.

Além disso, excluímos informações referentes a endereços com nível de granularidade alta (ruas, bairros, números de endereço) pois, apesar de adicionarem informações relativamente úteis, esse nível de detalhamento tem pouco impacto para nossas análises que foram feitas, em sua maioria, em nível de cidade. Dessa forma, também eliminamos dados sobre a instituição privada ser comunitária ou confessional, uma vez que nossa análise não se compromete a estudar especificamente estes tipos de instituições privadas.

Apesar das remoções, incluímos dados do IBGE sobre a quantidade de população existente e sobre a área em cada município, com intuito de tornar possível a filtragem de cidades que possuam determinada população e/ou área nas análises.

A partir disso, o recorte resultante é composto por informações de docentes, dados temporais, geográficos e identificadores da instituição.

A criação de um ambiente analítico para esse domínio pode possui diversos motivos, sendo a principal e mais provável deles a realização de consultas de dados da infraestrutura física e de funcionários da instituição com intuito de acompanhar e executar políticas públicas no segmento de educação. A partir disso, as análises envolvem comparações sobre docentes ou entre instituições de diferentes tipos (pública ou privada), localizadas em municípios e estados distintos. Por exemplo, são realizadas análises sobre a diversidade de funcionários ou sobre a concentração de instituições em municípios.

2. Modelagem

As únicas modificações realizadas foram as trocas de nomes dos atributos por conveniência.

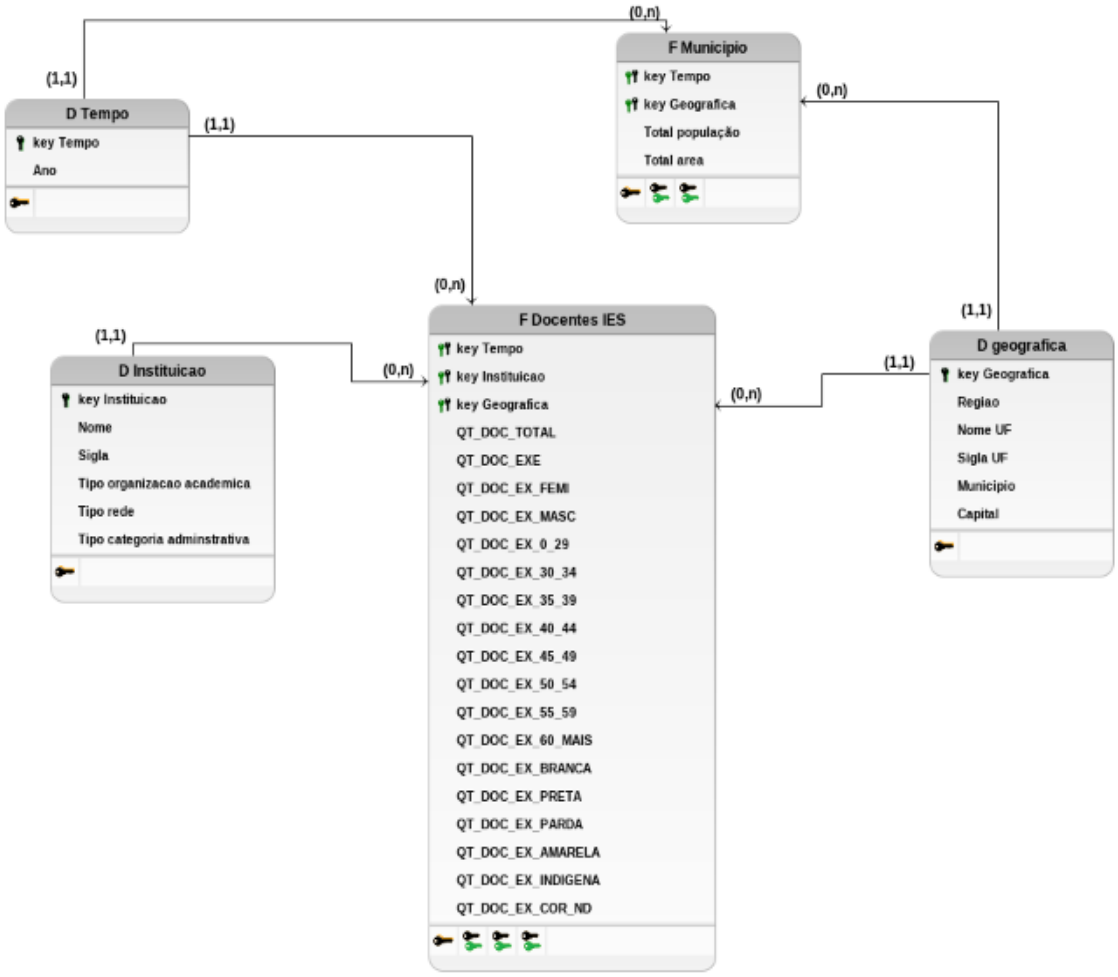

Figura 1. Modelo anterior

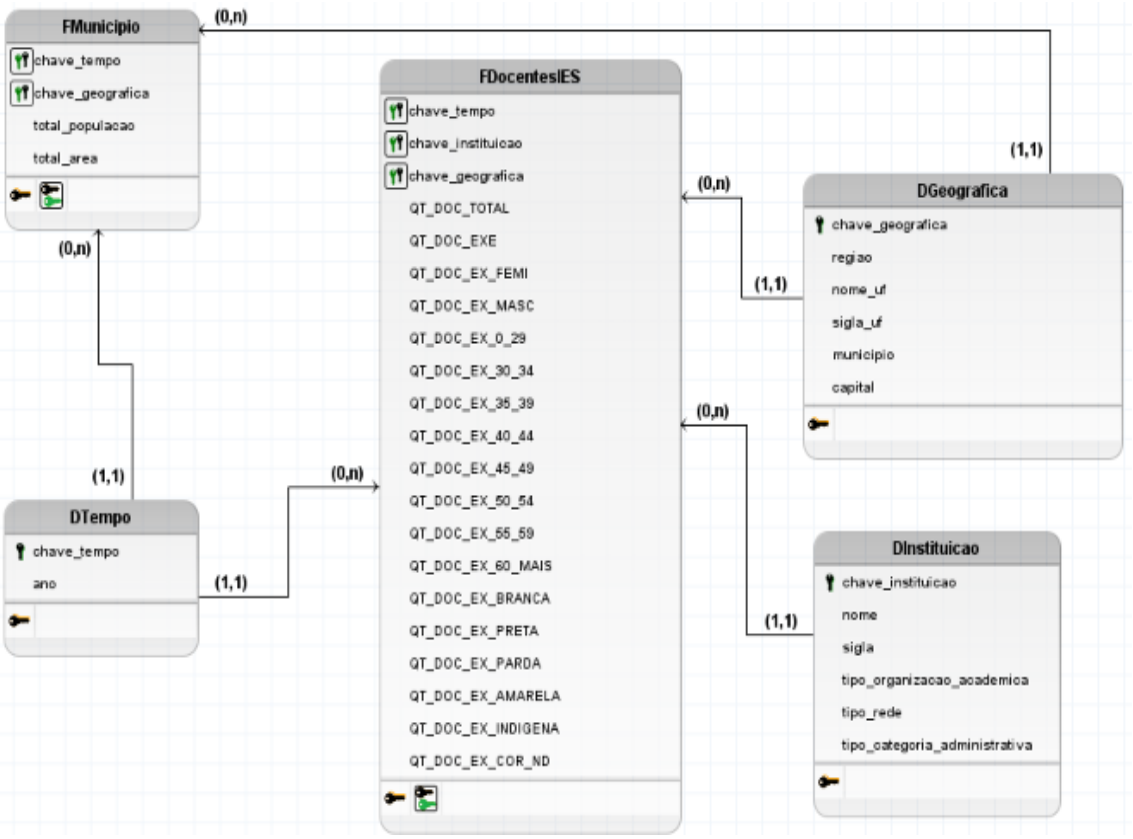

Figura 2. Modelo modificado

3. Dicionário de dados

3.1. DTempo

- **chave_tempo**: chave primária substituta da dimensão tempo.
- **ano**: ano de realização do censo.

3.2. DGeografica

- **chave_geografica**: chave primária substituta da dimensão geográfica.
- **regiao**: nome da região do país da sede administrativa ou reitoria.
- **nome_uf**: nome do Estado da sede administrativa ou reitoria.
- **sigla_uf**: sigla do Estado da sede administrativa ou reitoria.
- **municipio**: nome da cidade da sede administrativa ou reitoria.
- **capital**: flag que indica se a Cidade é capital do Estado.

3.3. DInstituicao

- **chave_instituicao**: chave primária substituta da dimensão instituição.
- **nome**: nome da Instituição.
- **sigla**: sigla da Instituição.
- **tipo_organizacao_academica**: informa o Tipo de Organização Acadêmica (Universidade, Faculdade, etc.).
- **tipo_rede**: informa a Rede de ensino (Pública ou Privada).
- **tipo_categoria_administrativa**: tipo de Categoria Administrativa.

3.4. FMunicipio

- **chave_tempo**: chave primária e estrangeira para a dimensão tempo.
- **chave_geografica**: chave primária e estrangeira para a dimensão geográfica.
- **total_populacao**: quantidade total de população no município.
- **total_area**: area total do município.

3.5. F Docentes IES

- **chave_tempo**: chave primária e estrangeira para a dimensão tempo.
- **chave_instituicao**: chave primária e estrangeira para a dimensão instituição.
- **chave_geografica**: chave primária e estrangeira para a dimensão geográfica.
- **QT_DOC_TOTAL**: quantidade total de docentes (em exercício e afastados).
- **QT_DOC_EXE**: quantidade total de docentes em exercício.
- **QT_DOC_EX_FEMI**: quantidade de docentes em exercício do sexo feminino.
- **QT_DOC_EX_MASC**: quantidade de docentes em exercício do sexo masculino.
- **QT_DOC_EX_0_29**: quantidade de docentes em exercício - até 29 anos.
- **QT_DOC_EX_30_34**: quantidade de docentes em exercício - de 30 a 34 anos.
- **QT_DOC_EX_35_39**: quantidade de docentes em exercício - de 35 a 39 anos.
- **QT_DOC_EX_40_44**: quantidade de docentes em exercício - de 40 a 44 anos.
- **QT_DOC_EX_45_49**: quantidade de docentes em exercício - de 45 a 49 anos.
- **QT_DOC_EX_50_54**: quantidade de docentes em exercício - de 50 a 54 anos.
- **QT_DOC_EX_55_59**: quantidade de docentes em exercício - de 55 a 59 anos.
- **QT_DOC_EX_60_MAIS**: quantidade de docentes em exercício - de 60 anos ou mais.
- **QT_DOC_EX_BRANCA**: quantidade de docentes em exercício - Cor/Raça branca.
- **QT_DOC_EX_PRETA**: quantidade de docentes em exercício - Cor/Raça preta.
- **QT_DOC_EX_PARDA**: quantidade de docentes em exercício - Cor/Raça parda.
- **QT_DOC_EX_AMARELA**: quantidade de docentes em exercício - Cor/Raça amarela.
- **QT_DOC_EX_INDIGENA**: quantidade de docentes em exercício - Cor/Raça indígena.
- **QT_DOC_EX_COR_ND**: quantidade de docentes em exercício - Cor/Raça não dispõe da informação ou não declarada.

4. ETL

4.1. Pentaho

Para o processo de ETL, inicialmente foi utilizado o Pentaho, mas por dificuldades com a ferramenta em relação à quantidade de dados utilizados, optamos por seguir com o Google Colab para o processo de transformação dos dados. Dessa forma, seguimos os mesmos passos já estruturados no Pentaho mostrados nos diagramas abaixo:

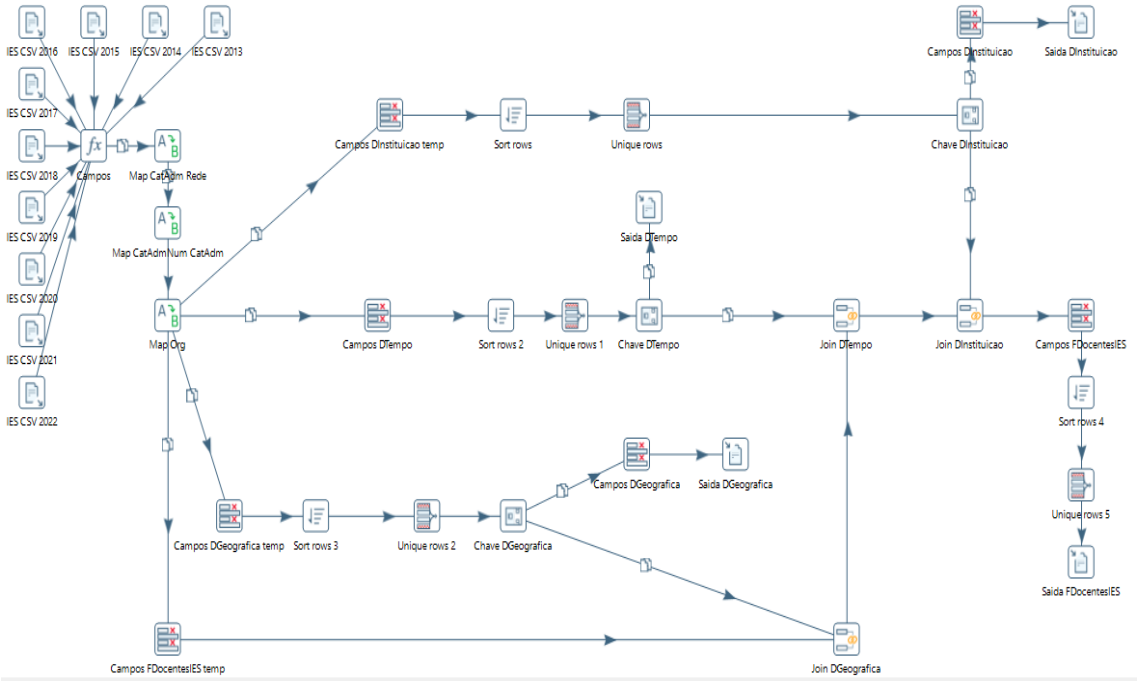

Figura 3. Diagrama das dimensões e do fato FDocentesIE

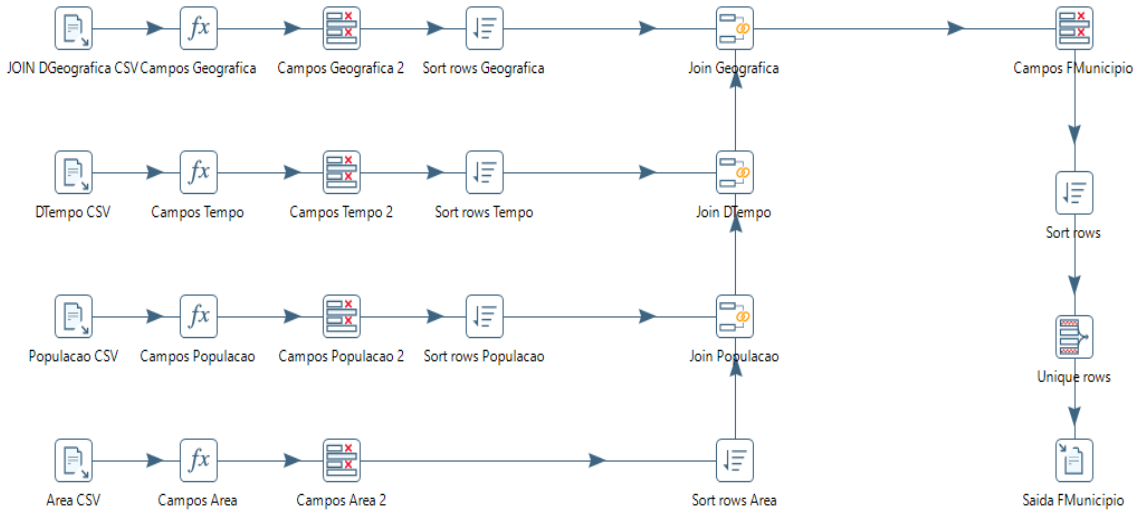

Figura 4. Diagrama do fato FMunicipio

4.2. Mapeamento

Base de Dados de Origem	Campo Origem	Tabela Destino	Campo Destino	Observação
IES	NU_ANO_CENSO	DTempo	ano	-
IES	NO_REGIAO_IES	DGeografica	regiao	-
IES	NO_UF_IES	DGeografica	nome_uf	-
IES	SG_UF_IES	DGeografica	sigla_uf	-
IES	NO_MUNICIPIO_IES	DGeografica	municipio	-
IES	IN_CAPITAL_IES	DGeografica	capital	-
IES	TP_ORGANIZACAO_ACADEMICA	DInstituicao	tipo_organizacao_academica	Mapeamento das flags para seus significados: 1. Universidade 2. Centro Universitário 3. Faculdade 4. Instituto Federal de Educação, Ciência e Tecnologia 5. Centro Federal de Educação Tecnológica
IES	TP_REDE	DInstituicao	tipo_rede	Mapeamento das flags para seus significados 1. Pública 2. Pública 3. Pública 4. Privada 5. Privada 6. Privada 7. Especial 8. Privada 9. Privada
IES	TP_CATEGORIA_ADMINISTRATIVA	DInstituicao	tipo_categoria_administrativa	Mapeamento das flags para seus significados: 1. Pública Federal 2. Pública Estadual 3. Pública Municipal 4. Privada com fins lucrativos 5. Privada sem fins lucrativos 6. Privada - Particular em sentido estrito 7. Especial 8. Privada comunitária 9. Privada confessional
IES	NO_IES	DInstituicao	nome	-
IES	SG_IES	DInstituicao	sigla	-
IES	QT_DOC_TOTAL	FDocentesIES	QT_DOC_TOTAL	-
IES	QT_DOC_EXE	FDocentesIES	QT_DOC_EXE	-
IES	QT_DOC_EX_FEMI	FDocentesIES	QT_DOC_EX_FEMI	-
IES	QT_DOC_EX_MASC	FDocentesIES	QT_DOC_EX_MASC	-
IES	QT_DOC_EX_0_29	FDocentesIES	QT_DOC_EX_0_29	-
IES	QT_DOC_EX_30_34	FDocentesIES	QT_DOC_EX_30_34	-
IES	QT_DOC_EX_35_39	FDocentesIES	QT_DOC_EX_35_39	-
IES	QT_DOC_EX_40_44	FDocentesIES	QT_DOC_EX_40_44	-
IES	QT_DOC_EX_45_49	FDocentesIES	QT_DOC_EX_45_49	-
IES	QT_DOC_EX_50_54	FDocentesIES	QT_DOC_EX_50_54	-
IES	QT_DOC_EX_55_59	FDocentesIES	QT_DOC_EX_55_59	-
IES	QT_DOC_EX_60_MAIS	FDocentesIES	QT_DOC_EX_60_MAIS	-
IES	QT_DOC_EX_BRANCA	FDocentesIES	QT_DOC_EX_BRANCA	-
IES	QT_DOC_EX_PRETA	FDocentesIES	QT_DOC_EX_PRETA	-
IES	QT_DOC_EX_PARDA	FDocentesIES	QT_DOC_EX_PARDA	-
IES	QT_DOC_EX_AMARELA	FDocentesIES	QT_DOC_EX_AMARELA	-
IES	QT_DOC_EX_INDIGENA	FDocentesIES	QT_DOC_EX_INDIGENA	-
IES	QT_DOC_EX_COR_ND	FDocentesIES	QT_DOC_EX_COR_ND	-
POPULACAO_MUN	populacao	FMunicipio	total_populacao	-
AREA_MUN	AREA_KM2	FMunicipio	total_area	-

Figura 5. Mapeamento dos dados

4.3. Google Colab - Diagrama das operações

No Google Colab, foram utilizadas a biblioteca pandas para a manipulação e leitura de dados em csv, e o SQLite para execução de consultas com junções de tabelas.

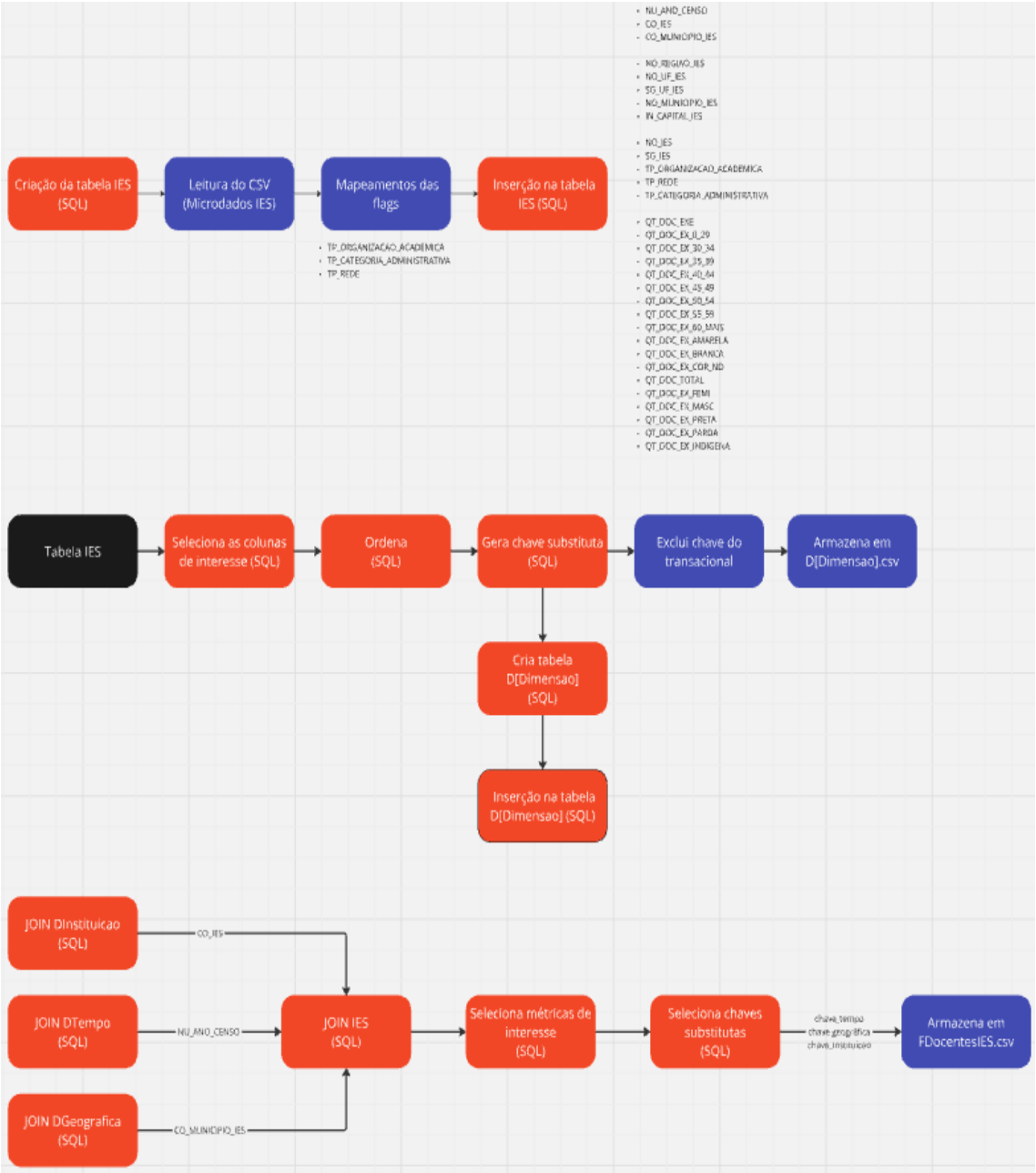

Figura 6. Processo de criação das dimensões (todas) e o fato FDocentesIES.

O diagrama acima representa o processo de criação das dimensões DTempo, DInstituicao, DGeografica e o fato FDocentesIES. Inicialmente, foi criada uma tabela IES no banco SQLite para armazenar os Microdados do INEP de todos os anos trabalhados (2013-2022), também realizando o mapeamento das colunas que estavam armazenadas em tags para campos em texto. Em seguida, para cada dimensão, foi executada uma consulta em que foram selecionadas suas colunas-alvo e também gerada uma chave substituta; o retorno desta consulta foi então armazenado em outras tabelas com o nome de cada dimensão, sendo também salvado em arquivos csv com a chave transacional removida. Por fim, as tabelas das dimensões foram utilizadas para relacionar as chaves substitutas com as métricas dos fatos, por meio de junções de tabelas com a tabela IES dos Microdados.

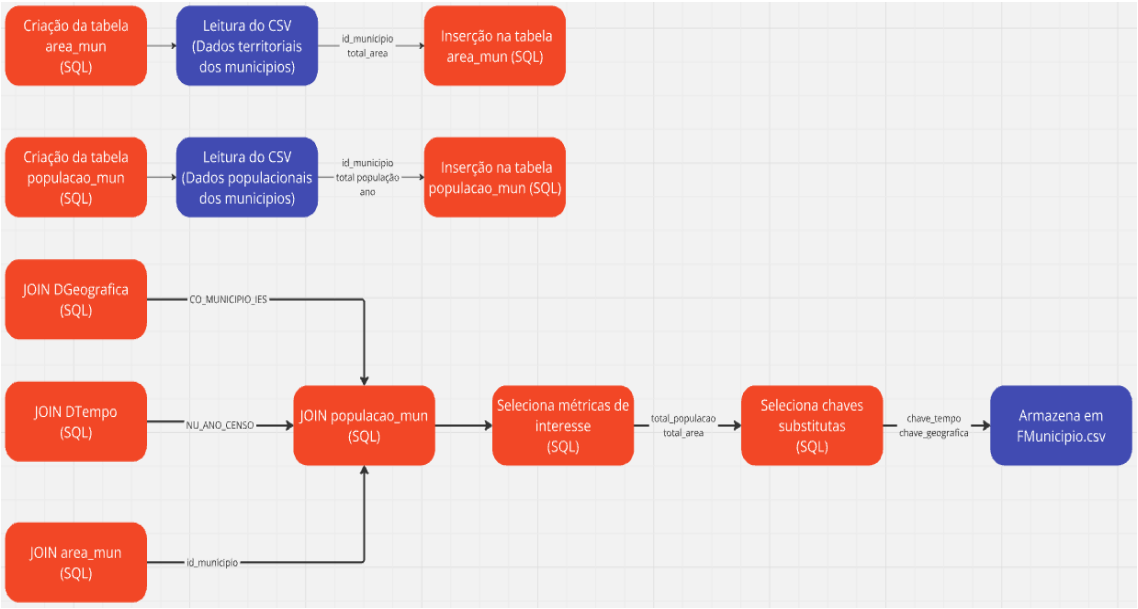

Figura 7. Processo criação do fato FMunicipio

Após a execução detalhada anteriormente, foi executado o processo criação do fato FMunicipio que contém dados de população e área dos municípios brasileiros. Esse processo começa com a leitura e armazenamento dos dados de população e área em tabelas SQL. Em seguida, também foi executada uma consulta em que houve uma junção de tabelas com as dimensões DGeografica e DTempo para relacionar as chaves substitutas com os fatos dos municípios e então salvar o retorno da consulta num arquivo csv.

5. Análises

5.1. Quantidade de IES dos municípios mais populosos em 2022

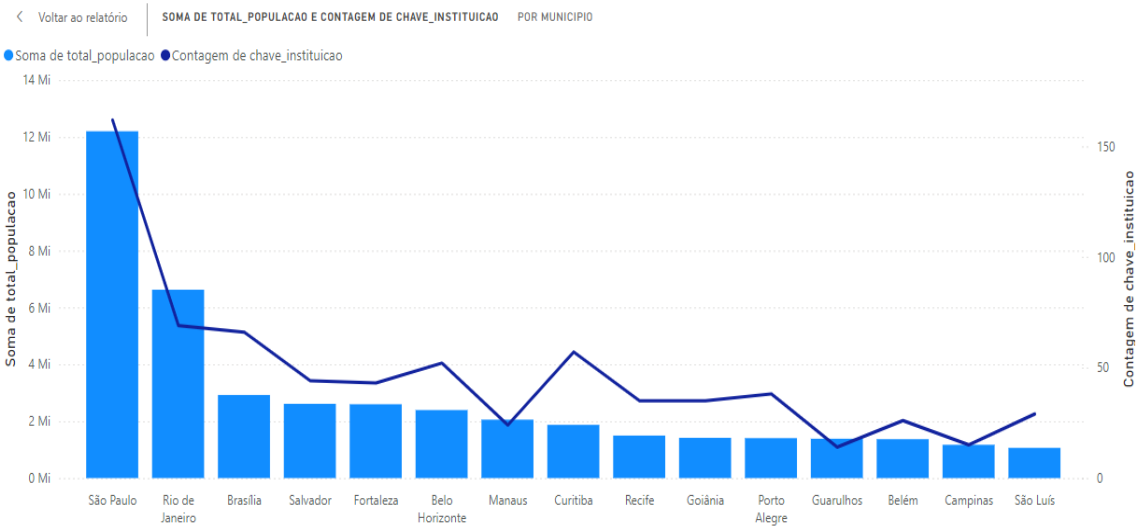

Figura 8. Gráfico de colunas agrupadas e linha

Observamos que a popularidade de um município e sua quantidade de instituições segue uma relação mais ou menos linear.

5.2. Quantidade de docentes dos municípios com maiores áreas em 2022

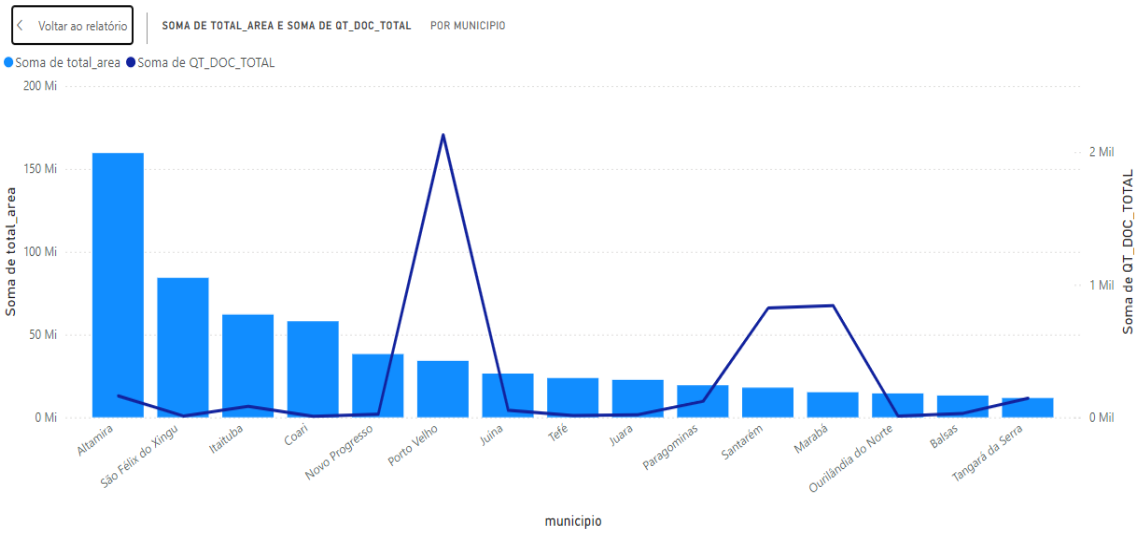

Figura 9. Gráfico de colunas agrupadas e linha

Diferentemente do gráfico anterior, não verificamos relação linear entre a área de um município e a quantidade de docentes.

5.3. Quantidade de docentes em exercício em instituições públicas e privadas ao longo do tempo

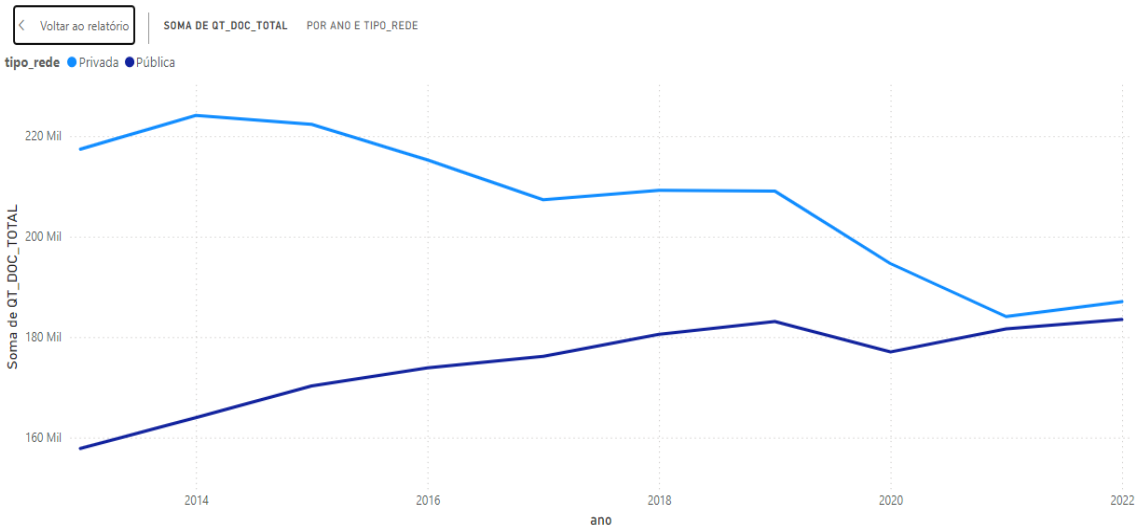

Figura 10. Gráfico de linhas

No inicío, há uma grande diferença entre docentes privados, mas com o passar do tempo essa diferença vai dimuindo até que as quantidades de cada categoria estejam equilibradas. A quantidade de docentes privados sempre é superior.

5.4. Quantidade de docentes brancos, pretos e pardos por Instituição de São Paulo ou do Rio de Janeiro em 2022

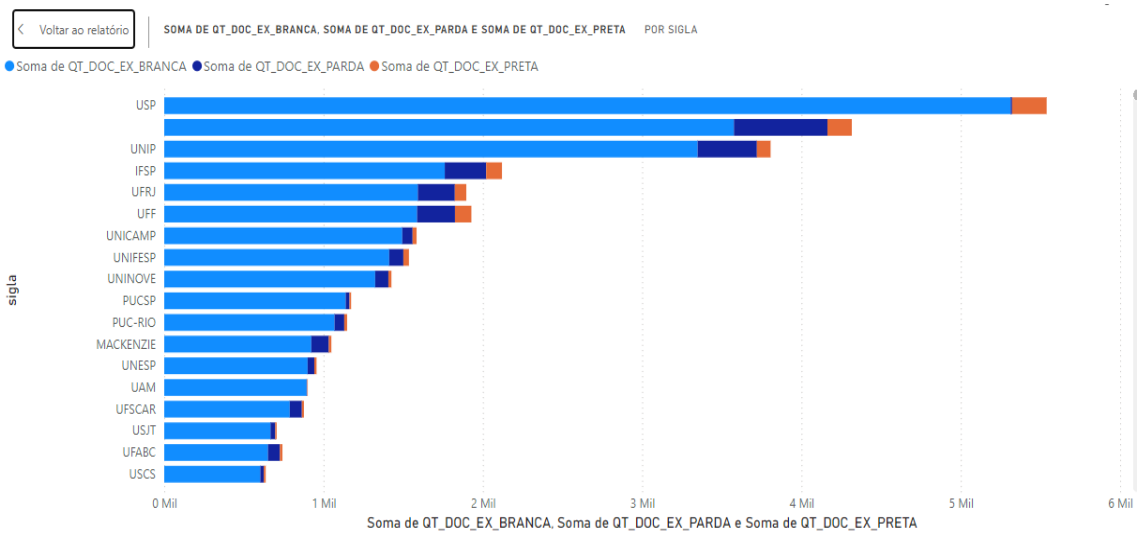

Figura 11. Gráfico de barras empilhadas

Em todas as instituições do gráfico, a quantidade de docentes brancos é muito mais numerosa que as outras categorias.Em segundo lugar de numerosidade temos docentes pardos e, em terceiro, docentes pretos.

5.5. Quantidade de docentes masculinos e femininos por cidade de 2013 a 2022

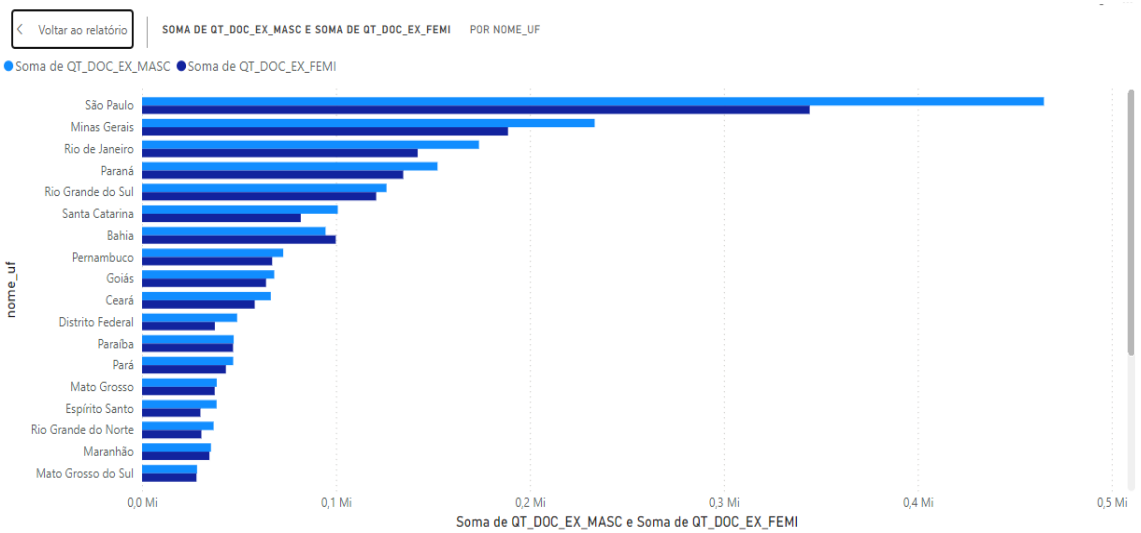

Figura 12. Gráfico de barras clusterizado

De maneira geral, as quantidades são equilibradas, sendo a quantidade de docentes masculinos um pouco maior. No gráfico, temos dois casos que fogem dessa regra geral: o primeiro é São Paulo, onde a quantidade de docentes masculinos é bem maior que a de femininos; o segundo caso é bahia, onde as mulheres são mais numerosas.

5.6. Quantidade de docentes em exercício por faixa etária nas universidades do Rio de Janeiro

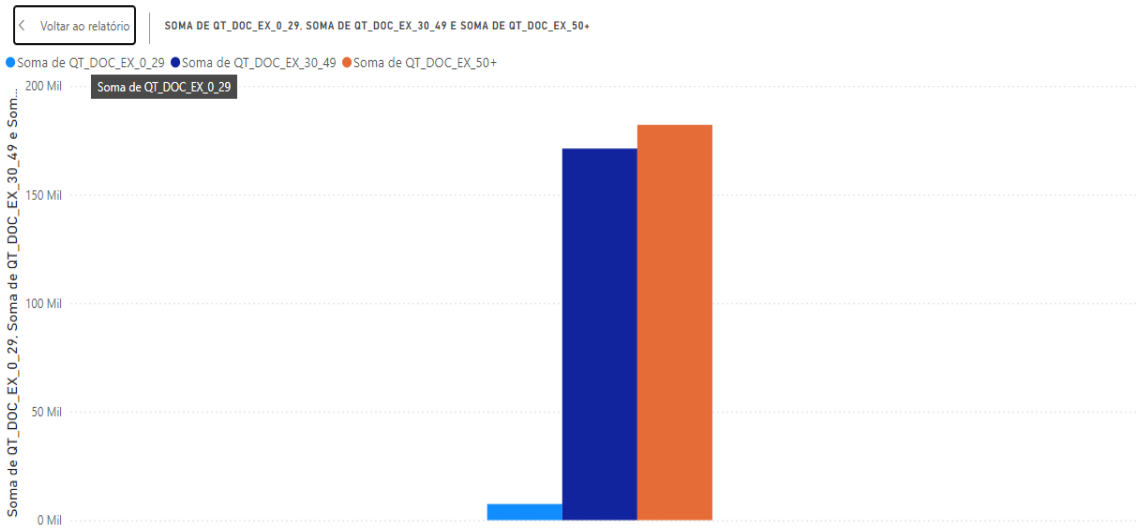

Figura 13. Gráfico de colunas clusterizado

Observamos que a faixa etária que vai até 29 anos é muito pouco numerosa em relação as outras duas categorias, as quais existem em quantidades equilibradas. Dessas outras duas, a faixa de docentes com 50 anos ou mais é um pouco mais numerosa que a faixa que vai dos 30 aos 49 anos.

5.7. Porcentagem de IES em capitais e fora de capitais em 2022

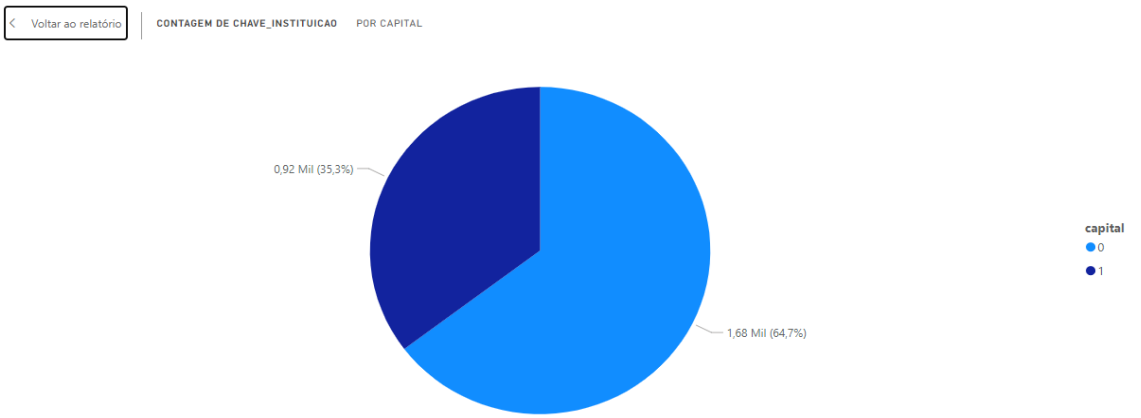

Figura 14. Gráfico de pizza

Verificamos que mais da metade das instituições não residem em uma capital.

5.8. Porcentagem de IES em cada região em 2022

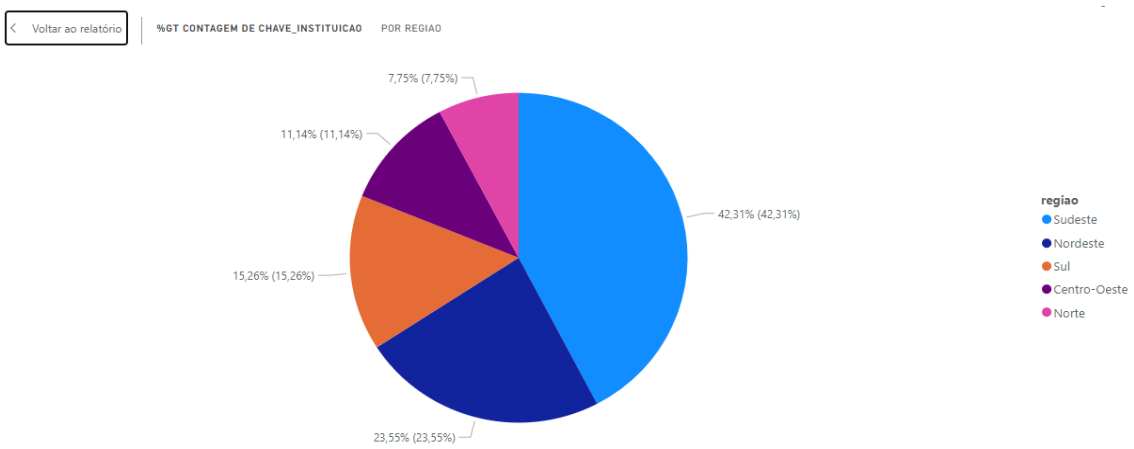

Figura 15. Gráfico de pizza

Estando em primeiro lugar, a região sudeste concentra pouco menos da metade as instituições do país. Sepois temos a região Nordeste, depois Sul, Centro-Oeste e Norte, nesta ordem.

5.9. Municípios com maiores quantidades de IES em 2022

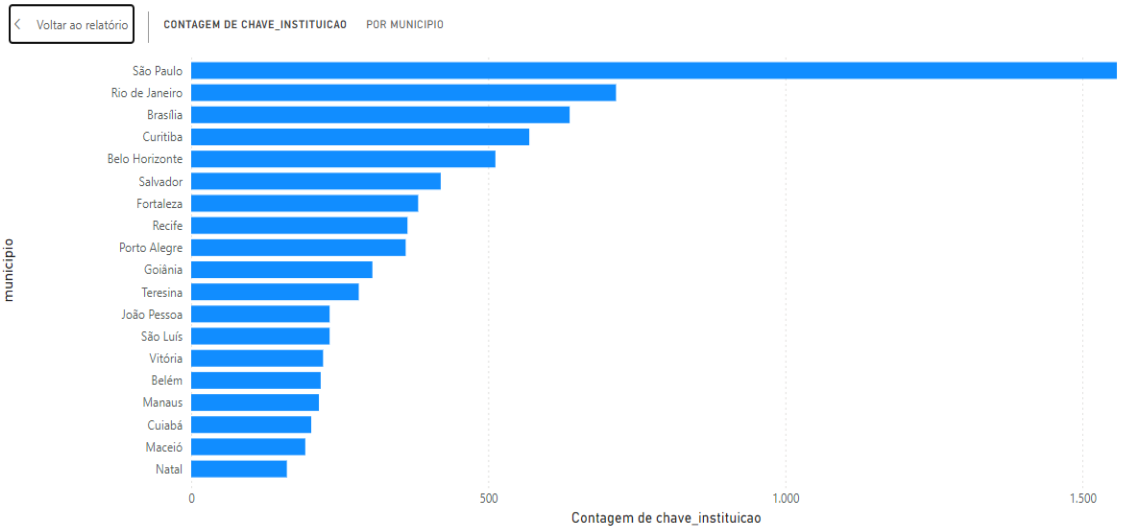

Figura 16. Gráfico de barras empilhadas

São Paulo lidera concentrando mais ou menos o dobro que o segundo lugar, Rio de Janeiro, a partir do qual, a quantidade diminui pouco a pouco a cada município.

6. Divisão do trabalho

- Daniel Arruda Ponte: Modelagem, seleção dos dados e processo de ETL.
- Matheus da Cruz Percine Pinto: Análises realizadas.
- Igor Santos: Modelagem e seleção dos dados.

7. Considerações finais

Consideramos o resultado do nosso ambiente analítico satisfatório, mas há pontos de melhoria como detalhes nas visualizações de dados como os nomes de variáveis mais explícitas para melhor entendimento das métricas apresentadas nos gráficos. Também há a possibilidade de explorar ferramentas mais avançadas que permitam maior flexibilidade para o usuário montar consultas próprias.

8. Referências

- Processo de ETL no Google Colab
- Board Workflow ETL no Miro