Belo Horizonte
2023
ALLAN DAYRELL MARCELINO
BRUNO SOARES E SILVA
EDUARDO HENRIQUE DOS SANTOS CORREIA
LUCAS ARAUJO DE MACEDO E SILVA
KAREN ROBERTA RODRIGUES DE SOUZA MOURA

TRABALHO PRÁTICO 2 DE BANCO DE DADOS:

Relatório

Trabalho interdisciplinar de Banco de Dados, do curso de Sistemas de Informação, Universidade Federal de Minas Gerais.

Orientador(a): Prof. Clodoveu Augusto Davis Junior

SUMÁRIO

1 FONTE DE DADOS

A escolha dos dados para a realização da tarefa proposta foi feita a partir dos dados fornecidos pela Prefeitura de Belo Horizonte. Visto que o site “dados.pbh.gov.br” fornece diversos tipos de dados que refletem a realidade da sociedade na cidade, escolhemos dois conjuntos que se relacionam e mostram, juntos, a situação de famílias de baixa renda em Belo Horizonte. Assim, conseguimos reunir dados acerca de famílias cadastradas no Cadastro Único de BH, que tem como objetivo a inclusão de programas de assistência social e redistribuição de renda; e de famílias de baixa renda presentes no Centro de Referência de Assistência Social (CRAS).

2 GITHUB

Repositório do Github:
https://github.com/edu-correia/TP2-IBD

Banco de Dados virtual (hospedado no Google Colab):
https://colab.research.google.com/drive/1UKUl-H_TRcS-Szs3OZZmeN1V0aTAG31G#scrollTo=nRcn4p8FjHRQ

3 ANÁLISE EXPLORATÓRIA

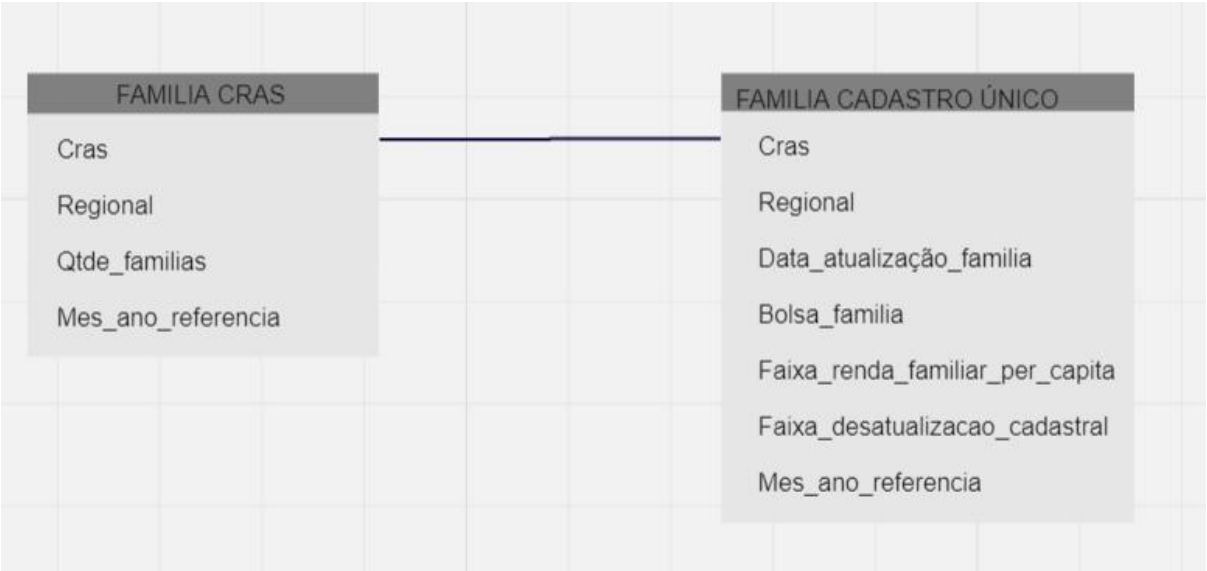
Preparação dos dados:

Com relação a preparação e limpeza dos dados, foram criadas métricas e filtros para garantir que não houvessem dados duplicados e que estes estivessem dispostos de maneira padronizada. Uma observação que pode ser feita é que quando foi feita a prescrição dos dados, foram utilizados caracteres que não são identificados como em “Endere�o FORA de Territ�rio TPSA”, e um erro de grafia no nome de uma das colunas (FAIXA_DESATUALICACAO_CADASTRAL), mas em ambos os casos não se trata de nada significante.

Definição de objetivos:

- Identificar o número total de famílias que recebem o benefício do Bolsa Família;
- Determinar o número total de pessoas cadastradas no CadÚnico;
- Contar o número total de Centros de Referência de Assistência Social (CRAS) distintos presentes no banco de dados;
- Contar quantos endereços no CadÚnico não possuem informações de georreferenciamento;
- Analisar a distribuição dos CRAS em diferentes regionais de Belo Horizonte;
- Verificar como a quantidade de pessoas cadastradas no CadÚnico está distribuída nas diferentes regionais da cidade;
- Analisar a distribuição de famílias com base nas faixas de renda familiar per capita no CadÚnico;
- Identificar e contar quantas famílias possuem desatualização cadastral no CadÚnico.

Diagrama UML:



Dicionário de Dados:

FAMÍLIA CRAS

COLUNA	DESCRIÇÃO	TIPO	RESTRIÇÃO
CRAS	Centro de Referência de Assistência Social no qual o endereço da família se encontra georreferenciado	Varchar(20)	PK
REGIONAL	Regional na qual o endereço da família se encontra georreferenciado	Varchar(20)	NOT NULL
QTDE_FAMILIAS	Número de famílias que se encontram georreferenciadas	Int	NOT NULL
MES_ANO_REFERENCIA	Mês e ano de referência dos dados	Date	NOT NULL

FAMÍLIA CADUNICO

COLUNA	DESCRIÇÃO	TIPO	RESTRIÇÃO
ID_FAMILIA_CADUNICO	Identificador da família no cadastro único	Int	PK
DATA_ATUALIZACAO_FAMILIA	Data da última atualização da família	Date	NOT NULL
BOLSA_FAMILIA	Recebe Programa BOLSA FAMÍLIA (SIM/NÃO) – Entre 10/2021	Varchar(20)	NOT NULL

	e 02/2023 Teve o nome de Auxilio Brasil		
FAIXA_RENDA_FAMILIAR_PER_CAPITA	Faixa da renda familiar per capita	Varchar(50)	NOT NULL
CRAS	Centro de Referência de Assistência Social no qual o endereço da família se encontra georreferenciado	Varchar(20)	FK(Família Cras)
REGIONAL	Regional na qual o endereço da família se encontra georreferenciado	Varchar(20)	NOT NULL
FAIXA_DESATUALIZACAO_CADASTRAL	Número de famílias que se encontram georreferenciadas	Varchar(20)	NOT NULL
MES_ANO_REFERENCIA	Mês e ano de referência dos dados	Date	NOT NULL

4 ANÁLISE DESCRITIVA

Quantas famílias recebem bolsa família:

```
[ ] q1 = "SELECT COUNT(*) AS qtd_bolsa_familia FROM cadunico WHERE BOLSA_FAMILIA = 'SIM';"
    runQuery(q1, "q1")
```

	qtd_bolsa_familia
0	137952

Quantas pessoas estão cadastradas no CadÚnico:

```
[ ] q1 = "SELECT SUM(QTDE_FAMILIAS) FROM cras;"
    runQuery(q1, "q1")
```

	SUM(QTDE_FAMILIAS)
0	295119

Quantidade de cras distintos no banco de daods:

```
[12] q1 = "SELECT COUNT(DISTINCT cras) AS cras_distintos FROM cadunico"
     runQuery(q1, "q1")
```

	cras_distintos
0	163

Quantos endereços não são georreferenciados:

```
[13] q1 = "SELECT COUNT(cras) AS ENDERECO_NAO_GEORREFERENCIADO FROM cadunico c WHERE c.cras= 'ENDERECO NAO GEORREFERENCIADO'"
     runQuery(q1, "q1")
```

	ENDERECO_NAO_GEORREFERENCIADO
0	15541

Quantas pessoas cadastradas por regional:

```
[14] q1 = "SELECT REGIONAL, COUNT(QTDE_FAMILIAS) FROM cras GROUP BY REGIONAL"
     runQuery(q1, "q1")
```

	REGIONAL	COUNT(QTDE_FAMILIAS)
0	BARREIRO	20
1	CENTRO-SUL	24
2	Endereco FORA Region	1
3	LESTE	20
4	NORDESTE	23
5	NOROESTE	23
6	NORTE	15
7	OESTE	21
8	PAMPULHA	18
9	VENDA NOVA	18

Quantos cras existem por regional:

```
[15] q1 = "SELECT REGIONAL, COUNT(*) AS qtd_cras FROM cras GROUP BY REGIONAL"
     runQuery(q1, "q1")
```

	REGIONAL	qtd_cras
0	BARREIRO	20
1	CENTRO-SUL	24
2	Endereco FORA Region	1
3	LESTE	20
4	NORDESTE	23
5	NOROESTE	23
6	NORTE	15
7	OESTE	21
8	PAMPULHA	18
9	VENDA NOVA	18

Quantas familias tem a faixa de desatualização cadastral de

- até 12 Meses
- 13 à 18 Meses
- 19 à 24 Meses
- 25 à 36 Meses
- 37 à 48 Meses
- acima de 48 Meses

```
[18] q1 = "SELECT FAIXA_DESATUALICACAO_CADASTRAL, COUNT(*) AS qtd_familias FROM cadunico GROUP BY FAIXA_DESATUALICACAO_CADASTRAL"
     runQuery(q1, "q1")
```

	FAIXA_DESATUALICACAO_CADASTRAL	qtd_familias
0	13 a 18 Meses	57681
1	19 a 24 Meses	34764
2	25 a 36 Meses	22091
3	37 a 48 Meses	17050
4	acima de 48 Meses	24497
5	ate 12 Meses	154577

Quantidade de pessoas com renda per capta até RS109.00 por regional

```
[19] q1 = "SELECT REGIONAL, COUNT(*) AS qtd_pessoas FROM cadunico WHERE FAIXA_RENDA_FAMILIAR_PER_CAPITA = 'Ate R$109.00' GROUP BY REGIONAL;"
     runQuery(q1, "q1")
```

	REGIONAL	qtd_pessoas
0	BARREIRO	15207
1	CENTRO SUL	11490
2	ENDERECO NAO GEORREFERENCIADO	4806
3	Endereco FORA Region	2
4	LESTE	12604
5	NORDESTE	13394
6	NOROESTE	11069
7	NORTE	20351
8	OESTE	10685
9	PAMPULHA	7122
10	VENDA NOVA	13743

Quantas pessoas cadastradas no CadÚnico por regional:

```
q1 = "SELECT REGIONAL, COUNT(*) AS qtd_pessoas FROM cadunico GROUP BY REGIONAL"
runQuery(q1, "q1")
```

	REGIONAL	qtd_pessoas
0	BARREIRO	41566
1	CENTRO SUL	25382
2	ENDERECO NAO GEORREFERENCIADO	10878
3	Endereco FORA Region	6
4	LESTE	30056
5	NORDESTE	36528
6	NOROESTE	30091
7	NORTE	44117
8	OESTE	30786
9	PAMPULHA	24390
10	VENDA NOVA	36860

Quantas familias tem a faixa de renda familiar per capta

Até RS109.00

RS109.01 até RS218.00

RS218.01 até 0.5 Salário Mínimo

Acima de 0.5 Salário Mínimo

```
q1 = "SELECT FAIXA_RENDA_FAMILIAR_PER_CAPITA, COUNT(*) AS qtd_familias FROM cadunico GROUP BY FAIXA_RENDA_FAMILIAR_PER_CAPITA"
runQuery(q1, "q1")
```

	FAIXA_RENDA_FAMILIAR_PER_CAPITA	qtd_familias
0	Acima de 0.5 Salario Minimo	103850
1	Ate R$109.00	120473
2	Entre R$109.01 ate R$218.00	22517
3	Entre R$218.01 ate 0.5 Salario Minimo	63820

Identificação de valores discrepantes:

Com base em nossa análise, constatamos que os dados pertinentes a este tópico envolvem a faixa de renda familiar per capita e a faixa de desatualização cadastral. Nesse contexto, observamos que não foram identificados valores críticos que se distanciam de maneira significativa do padrão geral dos dados.

Gráfico de faixa de desatualização:

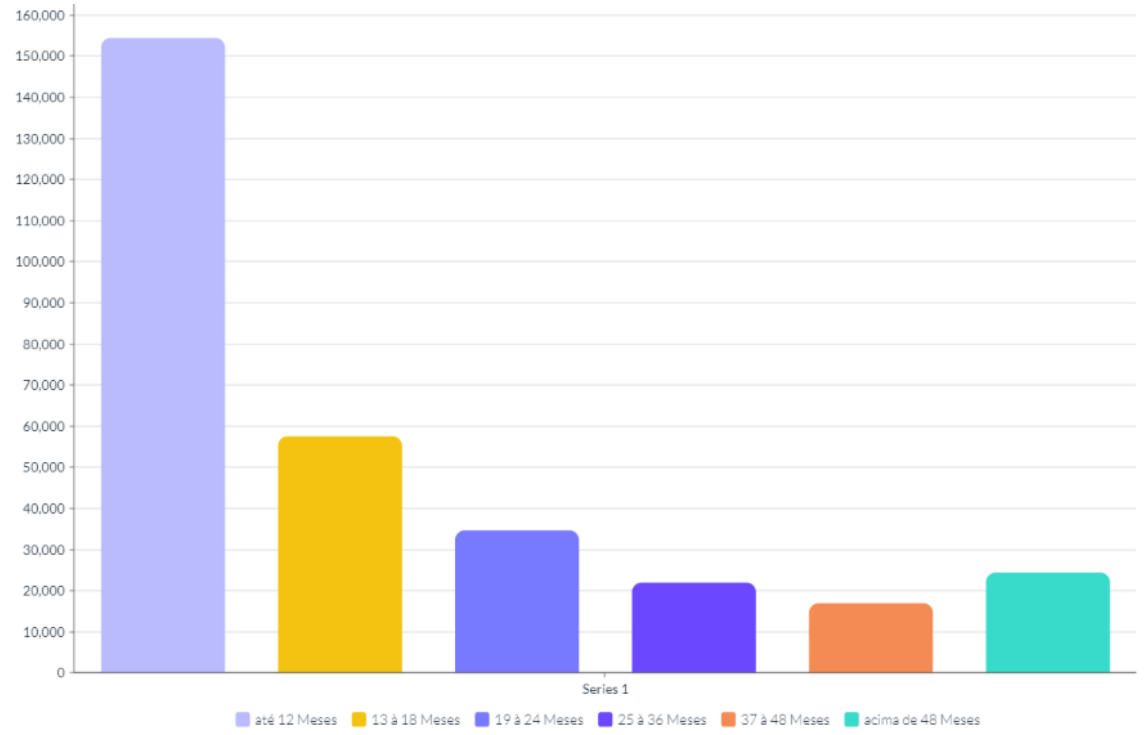
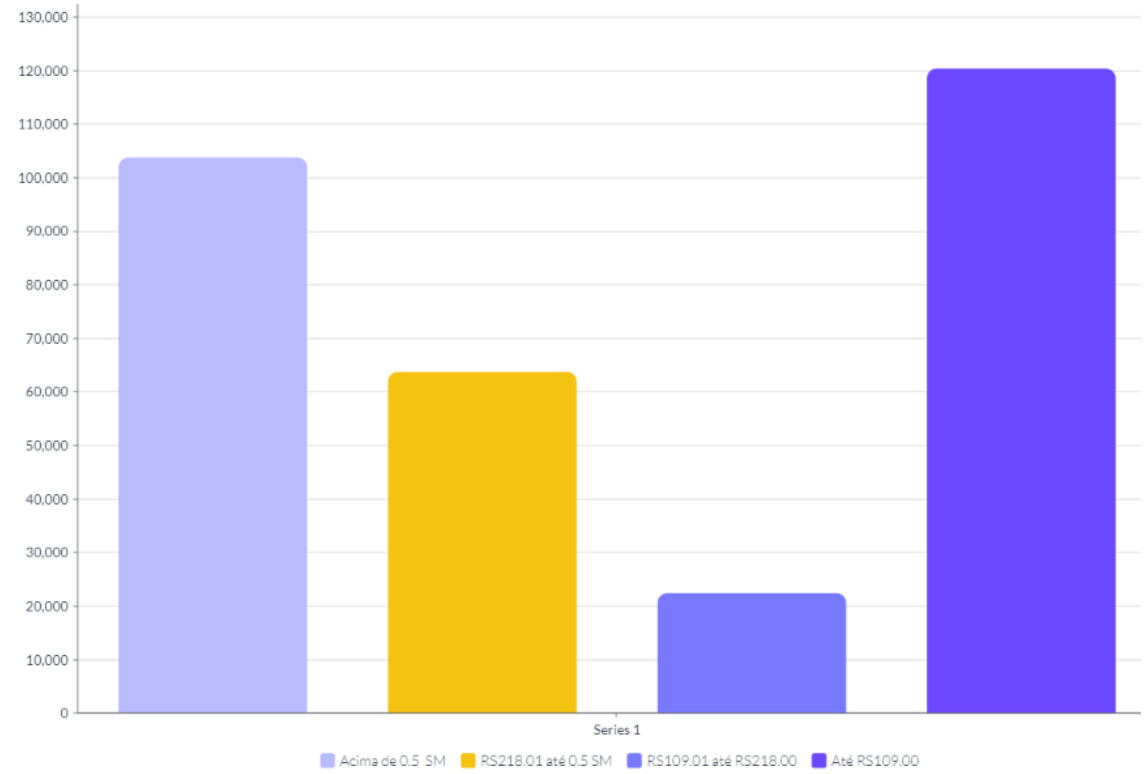


Gráfico de faixa de renda per capita:



Análise de correlação:

- Quantidade de famílias que recebem Bolsa Família vs. Quantidade de Pessoas Cadastradas:

  A quantidade de famílias que recebem o Bolsa Família pode estar positivamente correlacionada com a quantidade de pessoas cadastradas no CadÚnico. Essa associação positiva sugere que, à medida que o número de famílias beneficiárias do Bolsa Família aumenta, é esperado que o número total de pessoas cadastradas no CadÚnico aumente. Isso pode ocorrer porque o CadÚnico é utilizado para identificar e registrar informações sobre famílias que podem se beneficiar de programas sociais, como o Bolsa Família

- Quantidade de CRAS por regional vs. Quantidade de pessoas por regional:

  Ao observar a correlação positiva entre a quantidade de Centros de Referência de Assistência Social (CRAS) e a quantidade de pessoas em uma regional, podemos inferir que há uma possível relação entre a infraestrutura de assistência social e a densidade populacional. Em regiões onde há um aumento no número de CRAS, isso pode indicar um esforço para atender às necessidades de uma população mais numerosa. Esse fenômeno pode ser influenciado por vários fatores interligados.

- Quantidade de pessoas com renda per capita até 109 reais por regional vs. Quantidade de CRAS por regional:

  Uma correlação positiva sugere que, à medida que a quantidade de pessoas com renda per capita até 109 reais aumenta em uma regional, é esperado que a quantidade de CRAS também aumente nessa mesma regional. Isso pode refletir uma estratégia de distribuição de Centros de Referência de Assistência Social (CRAS) para atender às necessidades de áreas com uma maior concentração de pessoas em situação de vulnerabilidade socioeconômica.

Conclusões

Analisando os dados obtidos, foi possível observarmos uma série de fatores que só puderam ser notados graças às consultas e suas correlações, como por exemplo, a quantidade de pessoas que utilizam de serviços sociais como o CRAS e a bolsa família sendo cada vez maiores quando se trata de bairros mais populosos e

compostos por pessoas de baixa renda, sendo nestes bairros também onde se possuem o maior número de unidades de atendimento e sendo na prática os locais com maior uso do serviço. Outro fator interessante de se notar é a quantidade de endereços não georreferenciados, ainda mesmo nos dias atuais, o que demonstra ser um serviço que não recebe o investimento e a atenção necessárias, levando em conta o grande volume de pessoas que dependem deste tipo de atendimento.

Logo, este tipo de análise é importante pois ajuda a ter clareza sobre quais áreas mais necessitam de serviços deste tipo e onde priorizar as atenções para futuras tomadas de decisão, ainda mais neste tópico, em específico, pois se trata de uma questão pouco discutida atualmente, mas que faz muita diferença, principalmente para setores mais carentes da população.

5 ANÁLISE CRÍTICA

CRAS:

- A divergência no campo "MES_ANO_REFERENCIA" em diversas tabelas é evidente, com formatos distintos para representar o mesmo período. Por exemplo, Janeiro de 2019 é expresso como "012019", Fevereiro de 2019 como "Fev/2019", e Janeiro de 2020 como "01/01/2020".

- A discrepância na coluna "CRAS" durante o período de agosto de 2023 é notável, com algumas entradas denominadas como "Endereço FPRA de território TPSA" e outras como "Endereço fora de TPSA".

- A presença de diferentes formas de representação na coluna "REGIONAL," como "ENDEREÇO FORA AREA CRAS", “Endereço não georreferenciado”
- e "Endereço FORA de Área CRAS," para indicar a mesma informação.

- A presença de IDs que não estão associados a nenhuma região na coluna "REGIONAL" representa uma inconsistência nos dados que deve ser abordada. A ausência de informação sobre a região correspondente a determinados IDs pode impactar negativamente a análise e compreensão dos dados.

- A presença de campos adicionais, como "TEMPO_VIVE_NA_RUA," "CONTATO_PARENTE_FORA_RUA," "DATA_NASCIMENTO," "IDADE," "SEXO," "AUXILIO_BRASIL," "POP_RUA," "GRAU_INSTRUCAO," "COR_RACA," e "VAL_REMUNERACAO_MES_PASSADO" em uma tabela específica do período 11/2022, enquanto esses campos estão ausentes em outras tabelas de Famílias de Baixa Renda no CRAS.

- A presença de IDs em branco na coluna "CRAS," especificamente os IDs 17 e 19, pode indicar a ausência de informação sobre o Centro de Referência de Assistência Social associado a esses registros.

CadÚnico:

- A presença de IDs em branco na coluna "CRAS," e “REGIÃO” especificamente os IDs 18, 47, 57, 60,70,95 da tabela 04/2019 pode indicar a ausência de informação sobre o Centro de Referência de Assistência Social associado a esses registros.
- Presença de diferentes formas de representação na coluna "CRAS," como "ENDEREÇO FORA AREA CRAS", “Endereço não georreferenciado” e "Endereço FORA de Área CRAS," para indicar a mesma informação.
- A divergência no campo "MES_ANO_REFERENCIA" em diversas tabelas é evidente, com formatos distintos para representar o mesmo período. Por exemplo, Maio de 2019 é expresso como "MAI2019", Abril de 2019 como "ABR/2019".
- A divergência na estrutura da coluna "DATA_ATUALIZACAO_FAMILIA," com algumas tabelas representando a data e hora como "dd/mm/aaaa hh:mm:ss" e outras como "dd/mm/aaaa”.

Essa disparidade pode acarretar em confusões e erros durante a análise de dados, comprometendo a padronização essencial para a comparação entre conjuntos de dados. Além disso, essa inconsistência dificulta a realização de consultas e filtragens eficientes, impactando a precisão das análises. A falta de uniformidade nos formatos também torna desafiadora a documentação adequada dos dados, essencial para garantir compreensão e colaboração eficazes. Corrigir essa divergência é crucial para assegurar a integridade e confiabilidade dos dados ao longo do tempo.

6 ANÁLISE DE INTEGRAÇÃO DE DADOS

Durante esta seção, apresentamos uma análise integrada proveniente de múltiplas fontes de dados com o propósito de identificar correlações e relações entre conjuntos diversos. O nosso objetivo é fornecer uma visão mais ampla sobre as informações disponíveis. Abaixo, detalhamos os principais passos e resultados desse processo:

Iniciamos o processo compreendendo as características e o conteúdo de cada fonte de dados utilizada. Isso engloba a análise de "Famílias de Baixa Renda no CRAS", que oferece informações sobre famílias carentes cadastradas no Cadastro Único de Belo Horizonte, visando inclusão em programas sociais e redistribuição de renda. Além disso, incorporamos a fonte "Famílias no Cadastro Único", que abrange dados de famílias de baixa renda em âmbito nacional, independentemente de serem beneficiárias do Programa Bolsa Família. Ambas as fontes, originárias do Cadastro Único de Belo Horizonte, compartilham o objetivo de inclusão em programas sociais e redistribuição de renda. Durante esse processo inicial, identificamos minuciosamente as informações fornecidas por cada fonte e os contextos associados.

Posteriormente, identificamos chaves de ligação ou chaves primárias que desempenham um papel crucial como elementos comuns para estabelecer conexões entre dados oriundos de diferentes fontes. Essas chaves incluem a quantidade de famílias que recebem o Bolsa Família, em conjunto com a quantidade de pessoas cadastradas. Além disso, consideramos a quantidade de Centros de Referência de Assistência Social (CRAS) por regional em correlação com a quantidade de pessoas por regional. Por fim, incorporamos a quantidade de pessoas com renda per capita até 109 reais por regional, relacionada à quantidade de CRAS por regional. Esses elementos fundamentais proporcionam uma base sólida para a integração eficaz dos conjuntos de dados, possibilitando uma análise abrangente e interconectada.

Como parte do processo, normalizamos e padronizamos os dados para garantir uniformidade nos formatos, unidades e terminologias. Essa etapa é crucial para facilitar a comparação e integração dos dados, garantindo a consistência ao longo da análise.

Prosseguimos com uma análise exploratória de dados, visando identificar padrões, tendências e correlações entre os dados integrados. Utilizamos gráficos, estatísticas descritivas e outras técnicas de visualização para a incompreensão aprofundada dos conjuntos de dados combinados.

Este processo de análise integrada visa não apenas identificar relações entre dados, mas também proporcionar uma base sólida para a tomada de decisões informadas e estratégias futuras.

7 ORGANIZAÇÃO DE EQUIPE

Cada membro desempenhou um papel importante para a realização do trabalho prático.

Banco de Dados: Eduardo Henrique dos Santos Correia

- Responsável pela montagem do banco de dados, desde a concepção até a implementação.
- Responsável por entender os requisitos específicos do projeto.

Análise Exploratória: Lucas Araujo de Macedo e Silva

- Responsável pela montagem do dicionário de dados dos dados obtidos.
- Responsável por montar o esquema conceitual dos dados.

Análise Exploratória: Allan Dayrell Marcelino e Bruno Soares e Silva

- Responsáveis por fazer a recuperação dos dados.
- Responsáveis por definir os objetivos da análise.
- Responsáveis por identificar valores discrepantes.
- Responsáveis pela análise de correlação.

Análise Crítica: Karen Roberta Rodrigues de Souza Moura

- Responsáveis por fazer a interpretação e avaliação de conjunto de dados.
- Responsáveis por identificar limitações nos dados.
- Responsáveis por identificar dificuldades na padronização de dados.

Análise de Integração: Karen Roberta Rodrigues de Souza Moura

- Responsável por entender como os dados estão relacionados entre si.