



Universalizando o acesso a dados de qualidade

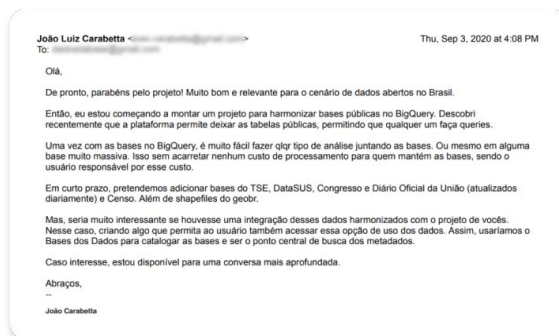
out/2023

A Base dos Dados é uma ONG com a missão de **universalizar o acesso a dados de qualidade.**



### De onde começamos

A BD nasce em 2019 a partir de um incômodo: como facilitar a análise de dados de forma simples e escalável?



### Primeiro data lake público do Brasil

Em 2020 resolvemos revolucionar o acesso de dados e construímos o 1º *data lake* público do Brasil, após muitas noites em claro e *code reviews*.



### Reconhecimento internacional

Em 2021 nosso projeto é premiado com o *Google Cloud Customer Awards* na categoria de Impacto Social.

## Nossos números atuais

- **21** *databasers* compõem a equipe da BD
- **+1,7 mil** pessoas em nossa comunidade
- **+35 mil** seguidores em redes sociais
- **+250 mil** usuários acessam a plataforma no Brasil e no mundo por ano
- **+4 milhões** de consultas aos dados realizadas desde 2019
- **+8 bilhões** de linhas de dados tratados e atualizados no data lake

# Data lake público

Ferramenta gratuita, acessível e centralizada para encontrar dados de qualidade

The image displays two overlapping screenshots of the Data Lake Público interface. The top screenshot shows a SQL query execution result with a table of data. The bottom screenshot shows a data catalog page for 'Relação Anual de Informações Sociais (RAIS)' with a search bar and a list of tables.

**Resultados da consulta**

id_municipio	id_municipio_ibge	id_municipio_ibge	id_municipio_ibge	nome	capital_uf	id_estado	id_regiao_saude	regiao_saude
1	1100023	110002	7	Araruama	0	1100023	11001	Vale do Jariari
2	1100106	110010	1	Guajará-Mirim	0	1100106	11004	Madeira-Mamoré
3	1100114	110011	15	Jariari	0	1100114	11003	Central
4	1100130	110013	39	Machadinho D'Oeste	0	1100130	11001	Vale do Jariari
5	1100205	110020	35	Porto Velho	1	1100205	11004	Madeira-Mamoré

**Relação Anual de Informações Sociais (RAIS)**

A Relação Anual de Informações Sociais (RAIS) é um relatório de informações socioeconômicas solicitado pela Secretaria de Trabalho do Ministério da Economia brasileiro às pessoas jurídicas e outros empregadores anualmente. Foi instituída pelo Decreto nº 11.000, de 1995.

Organização: Ministério da Economia (ME)

Cobertura temporal: 1995 - 2021

**Tema**

- ☐ Economia (183)
- ☐ Saúde (150)
- ☒ Meio Ambiente (129)
- ☐ Segurança, Crime, Violência e Conflito (129)
- ☐ Política (105)
- ☐ Educação (88)

- **+2TB de dados disponíveis ao público**
- **+600 tabelas de diferentes fontes**, como IBGE, Ministério da Economia, Educação, Tesouro, DataSUS, ANS, Banco Central, CVM
- **Buscador** para encontrar dados com facilidade
- Acesso via *data lake* online, download ou pacotes de programação (em [Python](#), [R](#), [Stata](#))

# Base dos Dados e BigQuery



Google Cloud



- A Base dos Dados hospeda seu *datalake* público no Google Cloud. O BigQuery é uma **plataforma que suporta consultas usando SQL**, ou seja, é possível usá-lo para consultar diretamente os dados da BD.
- O Google Cloud e o BigQuery são apenas ferramentas que a Base dos Dados utiliza para disponibilizar seus dados. Portanto, **a utilização dos dados da BD no BigQuery é feita através da sua conta Google**, por onde você controla sua cota de armazenamento, processamento e outras.



O *datalake* que você já conhece  
ainda **maior** e **especializado** para  
dados de seu interesse.



Consultoria para **fortalecer**  
**organizações que precisam**  
**tomar melhores decisões** com  
dados de qualidade.



Cursos para **empoderar**  
**pessoas no mundo dos**  
**dados.**

# Letramento em análise de dados

Uso das redes sociais e visualização de dados para mostrar **as melhores formas de usar o *datalake* público**



## Produção de análises e visualizações

Centenas análises com código aberto e reproduzível no [Github](#), divulgadas nas redes e [newsletter](#).



## Workshops e análises ao vivo

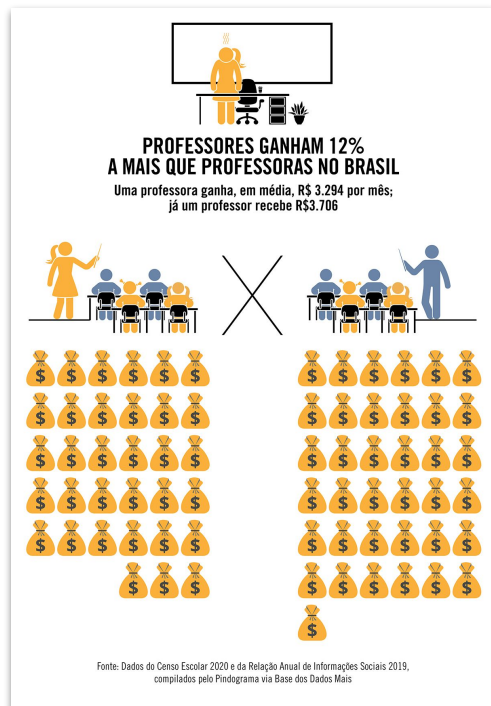
Conteúdos gravados de dicas e tutoriais com analistas da BD no [Youtube](#).



## Suporte para comunidade

Além de fornecer a matéria prima, prestamos suporte via Discord, Whatsapp, Telegram e redes sociais

# Professores e professoras têm remuneração igual no Brasil?



Na análise, a Piauí agrega **dados de docentes do Censo Escolar com a remuneração anual da RAIS**, para entender as desigualdades salariais entre homens e mulheres de diferentes regiões do país.

*Piauí =igualdades: Elas na sala de aula*



# Quantas hospitalizações podem ser evitadas com Atenção Primária?

Mapa da Taxa de Hospitalizações por Condições Sensíveis à Atenção Primária por Município em (%) por 2010

Fonte - IEPS



Christian Basilio Oliveira • 2º

Data Analyst | Data Journalist

6 m • Editado •

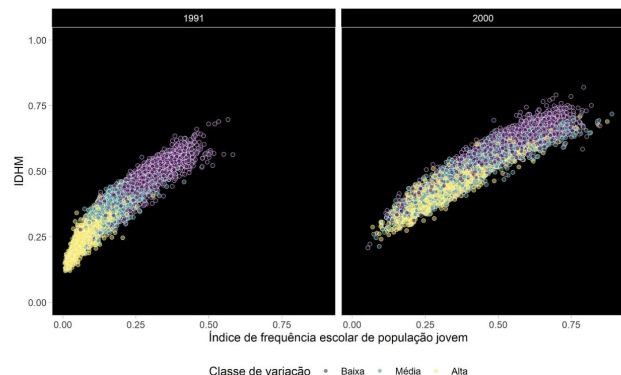
Recentemente estive analisando dados do [Instituto de Estudos para Políticas de Saúde \(IEPS\)](#) com dados disponibilizados pela [Base dos Dados](#) por meio de API. Com o auxílio do Python, pude realizar uma análise com o objetivo de observar o fator de hospitalizações por condições sensíveis à atenção primária ou básica no Brasil.

Christian, pesquisador da Fiocruz, compilou mais de 10 anos de dados de internações hospitalares do IEPS Data ([disponíveis na BD](#)) para **identificar áreas que necessitam de maior atenção e investimentos e analisar a efetividade da Política Nacional de Atenção Básica (PNAB).**

[Análise do Christian no LinkedIn](#)

## Cidades com maior IDH têm maior frequência escolar?

Four indicators are used to calculate the school attendance of the young population: percentages of 5 to 6-year-olds attending school, of 11 to 13-year-olds following the final years of elementary school, of 15 to 17-year-olds with complete elementary school, and of 18 to 20-year-olds with entire high school. See below how this indicator divides the three classes of variations in the MHDl between 1991 and 2000.



Influence of school attendance in MHDl variation between 1991 and 2000. Image by the author

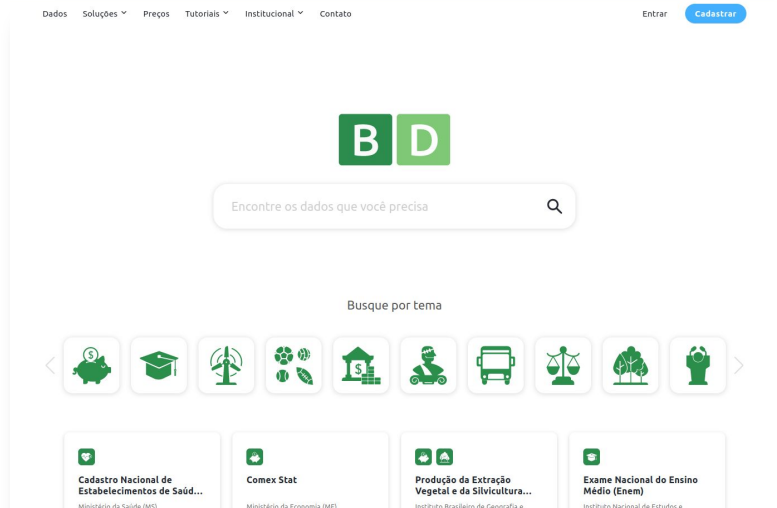
Fernando Barbalho (Tesouro Nacional) cruzou dados de frequência escolar e do Índice de Desenvolvimento Humano Municipal (IDHM) para **compreender o impacto da educação no desenvolvimento brasileiro**.

A série histórica completa, desde 1991, está disponível na base do [Atlas de Desenvolvimento Humano \(ADH\)](#) na BD.

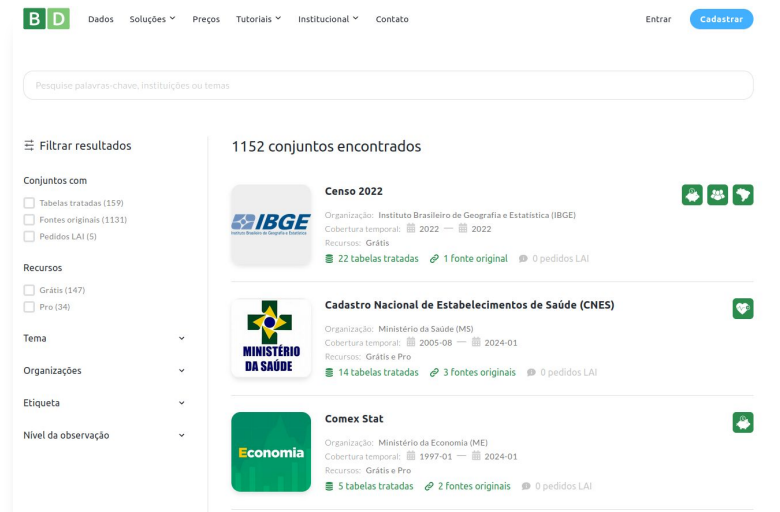
[\*Towards Data Science: An analysis of Brazilian census data\*](#)

# Como navegar pela Base dos Dados?

## Use a busca ou navegue pelos temas



## Filtre e explore



# Como navegar pela Base dos Dados?

## Conheça as tabelas

B

D

Dados

Soluções

Preços


Tutoriais

Institucional

Contato

Entrar

Cadastrar



### Sistema Nacional de Informações sobre Saneamento (SNIS)

Criado em 1996, o SNIS é uma unidade vinculada à Secretaria Nacional de Saneamento (SNS) do Ministério do Desenvolvimento Regional (MDR). Com abrangência nacional, reúne informações de caráter institucional, administrativo, operacional, ger...

[Ler mais >>](#)

#### Organização

Ministério do Desenvolvimento Regional (MDR)

#### Cobertura temporal

1995 - 2021

Dados

Tabelas tratadas

Serviços de Água e Esgoto nos Municípios

Prestadores de Água e Esgoto

Fontes originais

Série Histórica Municípios

Diagnóstico Anual de Água e Esgotos

Diagnóstico Anual de Resíduos Sólidos

### Serviços de Água e Esgoto nos Municípios

#### Consulta aos dados

SQL Python R Stata Download

Copie o código abaixo, [clique aqui](#) para ir ao *datalake* no BigQuery e cole no Editor de Consultas:

```
SELECT * FROM `basedosdados.br_mdr_snis.municipio_agua_esgoto` LIMIT 100
```

Copiar

Para usar o BigQuery basta ter uma conta Google. Primeira vez? [Siga o passo a passo.](#)

#### Descrição

Esta tabela contém informações e indicadores consolidados por município dos serviços de água e esgoto fornecendo um panorama detalhado do histórico e do estado da arte do saneamento básico no Brasil. Atenção! este dataset não representa o conjunto total de dados disponíveis no SNIS. Para detalhes sobre como fizemos a seleção dos dados na fonte original basta realizar o download dos arquivos auxiliares no final da página.

## Entenda as colunas

B

D

Dados

Soluções

Preços

Tutoriais

Institucional

Contato

Entrar

Cadastrar

### Cobertura temporal

1995 - 2021

### Colunas

Nome	Tipo No BigQuery	Descrição
ano	INT64	Ano
id_municipio	STRING	ID Município - IBGE 7 Dígitos
sigla_uf	STRING	Sigla da Unidade da Federação
populacao_atendida_agua	INT64	AG001 - População total atendida com abastecir
populacao_atendida_esgoto	INT64	ES001 - População total atendida com esgotame
populacao_urbana	INT64	População urbana do município do ano de referê
populacao_urbana_residente_agua	INT64	G06A - População urbana residente do(s) municí
populacao_urbana_atendida_agua	INT64	AG026 - População urbana atendida com abaste
populacao_urbana_atendida_agua_ibge	INT64	G12A - População total residente do(s) municí

### Nível da observação

Entidade	Colunas Correspondentes
Município	id_municipio
Ano	ano

### Informações adicionais

# Usando os dicionários

## Como ele se estrutura?

id_tabela	coluna	chave	cobertura_temporal	valor
microdados	local_nascimento	1	(1)	Hospital
microdados	local_nascimento	2	(1)	Outros estabelecimentos
microdados	local_nascimento	3	(1)	Domicílio
microdados	local_nascimento	4	(1)	Outros
microdados	local_nascimento	9	(1)	Ignorado
microdados	sexo	0	(1)	Ignorado
microdados	sexo	1	(1)	Masculino
microdados	sexo	2	(1)	Feminino
microdados	raca_cor	1	(1)	Branca
microdados	raca_cor	2	(1)	Preta
microdados	raca_cor	3	(1)	Amarela
microdados	raca_cor	4	(1)	Parda
microdados	raca_cor	5	(1)	Indígena
microdados	apgar1	99	(1)	Ignorado
microdados	apgar5	99	(1)	Ignorado
microdados	id_anomalia	1	1997, 1999(1)	Sim
microdados	id_anomalia	2	1997, 1999(1)	Não
microdados	id_anomalia	9	1997, 1999(1)	Ignorado
microdados	semana_gestacao_eb	1	2010(1)	Exame físico
microdados	semana_gestacao_eb	2	(1)	Outro método
microdados	semana_gestacao_eb	9	(1)	Ignorado
microdados	gestacao_agr	1	(1)	Menos de 22 semanas
microdados	gestacao_agr	2	(1)	22 a 27 semanas

## Como acessar?

### Dicionário

### Consulta aos dados

[SQL](#)[Python](#)[R](#)[Stata](#)[Download](#)

Estes dados estão disponíveis porque diversas pessoas colaboraram. Antes de baixar os dados, apoie você também com uma doação!

[Download dos dados](#)

# Diretórios da BD

O perfil completo de unidades como município, escola, UF, e mais.

```
SELECT
  t1.ano,
  t2.nome_uf AS estado,
  t2.nome AS municipio,
  t1.tempo_medio_deslocamento
FROM `basedosdados.br_mobilidados_indicadores.tempo_deslocamento_casa_trabalho` AS t1
JOIN `basedosdados.br_bd_diretorios_brasil.municipio` AS t2
ON t1.id_municipio = t2.id_municipio
```

Deixa mais prático o cruzamento de diferentes conjuntos criando relações entre entidades, como UF, município, escola, distrito, setor censitário e mais.

ano	tempo_medio_deslocamento
2010	28
2010	26
2010	41

# Vamos fazer juntas

- Acessar o pacote da base dos dados via R
- Responder a pergunta:

Quantas mulheres não fizeram o pré-natal adequado em Curitiba?

- Qual a base?
- Qual a cobertura temporal?
- Quais colunas eu vou precisar para responder essa pergunta?
- Preciso usar o dicionário ou o diretório?

# Como acessar a base dos dados no R?

- Criar projeto no BQ

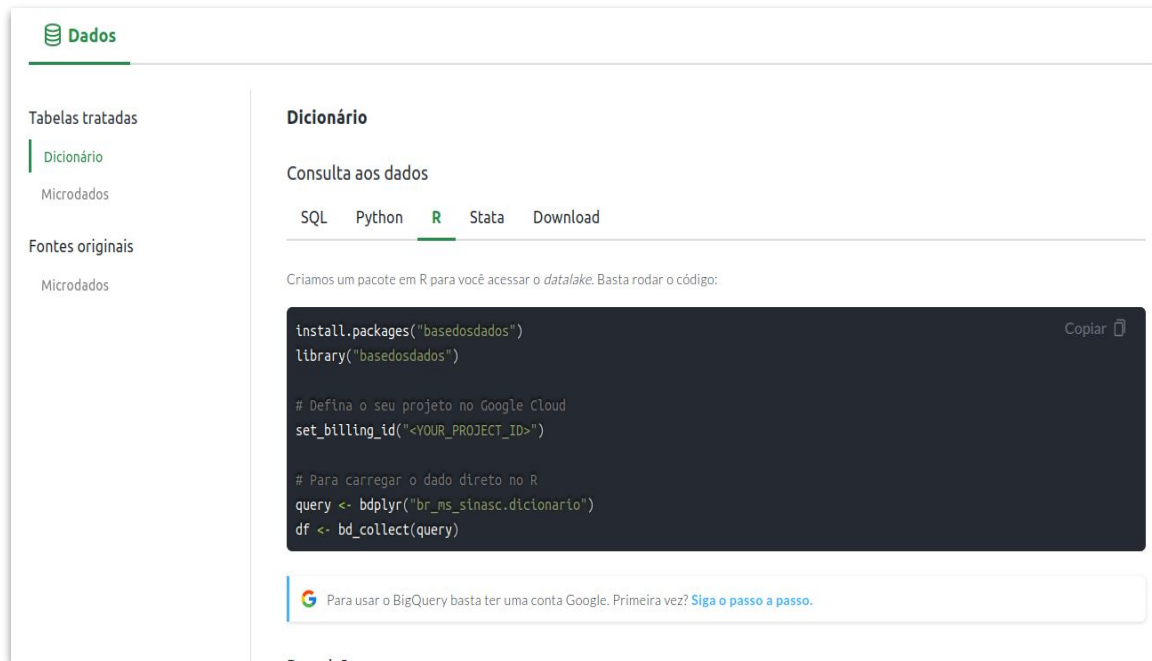
Para criar um projeto no Google Cloud basta ter um email cadastrado no Google. É necessário ter um projeto seu, mesmo que vazio, para você fazer queries em nosso *datalake* público.

1. [Acesse o Google Cloud](#). Caso for a sua primeira vez, aceite o Termo de Serviços.
2. **Clique em Create Project/Criar Projeto**. Escolha um nome bacana para o projeto.
3. **Clique em Create/Criar**



# Como acessar a base dos dados no R?

- Criar projeto no BQ
- Instalar pacote
- Definir o billing id
- Fazer sua primeira requisição



The screenshot shows the 'Dados' (Data) interface. On the left, there is a sidebar with a menu containing 'Tabelas tratadas', 'Dicionário', 'Microdados', 'Fontes originais', and 'Microdados'. The 'Dicionário' (Dictionary) section is selected. The main area is titled 'Dicionário' and has a sub-header 'Consulta aos dados'. Below this, there are tabs for 'SQL', 'Python', 'R', 'Stata', and 'Download'. The 'R' tab is active. The content area displays the following R code:

```
install.packages("basedosdados")
library("basedosdados")

# Define o seu projeto no Google Cloud
set_billing_id("<YOUR_PROJECT_ID>")





# Para carregar o dado direto no R
query <- bdplyr("br_ns_sinasc.dicionario")
df <- bd_collect(query)
```

Below the code, there is a note: 'Criamos um pacote em R para você acessar o *datalake*. Basta rodar o código:'. At the bottom, there is a Google logo and a link: 'Para usar o BigQuery basta ter uma conta Google. Primeira vez? [Siga o passo a passo.](#)'.

# Como acessar a base dos dados no R?

- Criar projeto no BQ
- Instalar pacote
- Definir o billing id
- Fazer sua primeira requisição
- Marcar todas as opções

Selecione o que o app **Tidyverse API Packages** pode acessar

	Associar suas informações pessoais a você no Google	<input checked="" type="checkbox"/>
	Ver o endereço de e-mail principal da sua Conta do Google	<input checked="" type="checkbox"/>
	Ver, editar, configurar e excluir seus dados do Google Cloud e ver o endereço de e-mail da sua Conta do Google.. <a href="#">Saiba mais</a>	<input type="checkbox"/>
	View and manage your data in Google BigQuery and see the email address for your Google Account. <a href="#">Saiba mais</a>	<input type="checkbox"/>

Confirme se o app Tidyverse API Packages é confiável

**ATENÇÃO!**

É preciso marcar inclusive as "caixinhas" que aparecem como opcionais

Essa é uma das principais causas de problemas com o R

# Como acessar a base dos dados no R?

- Criar projeto no BQ
- Instalar pacote
- Definir o billing id
- Fazer sua primeira requisição
- Marcar todas as opções

**Prontinho!**

Você está pronto para acessar as bases da BD com o R!

Existem 2 funções principais para puxar as informações

- `bdplyer`
- `read_sql`

# Vamos fazer juntas

- Responder a pergunta:

Quantas mulheres não fizeram o pré-natal adequado em Curitiba?

- Qual a base?
- Qual a cobertura temporal?
- Quais colunas eu vou precisar para responder essa pergunta?
- Preciso usar o dicionário ou o diretório?

# Exemplo

- Quais as características das mulheres não fizeram o pré-natal adequado em Curitiba?
  - Qual a base?

SINASC - Sistema de Nascidos Vivos
  - Qual a cobertura temporal?
  - Quais colunas eu vou precisar para responder essa pergunta?
  - Preciso usar o dicionário ou o diretório?

# Exemplo

- Quais as características das mulheres não fizeram o pré-natal adequado em Curitiba?
  - Qual a base?

SINASC - Sistema de Nascidos Vivos
  - Qual a cobertura temporal?

Vou escolher o ano mais recente para iniciar a análise
  - Quais colunas eu vou precisar para responder essa pergunta?
  - Preciso usar o dicionário ou o diretório?

# Exemplo

- Quais as características das mulheres não fizeram o pré-natal adequado em Curitiba?
  - Qual a base?

SINASC - Sistema de Nascidos Vivos
  - Qual a cobertura temporal?

Vou escolher o ano mais recente para iniciar a análise
  - Quais colunas eu vou precisar para responder essa pergunta?

`ano, id_municipio, classificacao_pre_natal, raca_cor, idade_mae, local_nascimento,`  
`escolaridade_mae, estado_civil_mae`
  - Preciso usar o dicionário ou o diretório?

# Exemplo

- Quais as características das mulheres não fizeram o pré-natal adequado em Curitiba?

- Qual a base?

SINASC - Sistema de Nascidos Vivos

- Qual a cobertura temporal?

Vou escolher o ano mais recente para iniciar a análise

- Quais colunas eu vou precisar para responder essa pergunta?

```
ano, id_municipio, classificacao_pre_natal, raca_cor, idade_mae,  
escolaridade_mae, estado_civil_mae
```

- Preciso usar o dicionário ou o diretório?

Sim, dicionário para saber o que usar como pré natal adequado e o diretório para puxar o município de curitiba



# Exemplo

```
SELECT id_municipio
FROM `basedosdados.br_bd_diretorios_brasil.municipio`
WHERE nome = 'Curitiba'
```

```
SELECT  raca_cor, idade_mae, escolaridade_mae, estado_civil_mae
FROM `basedosdados.br_ms_sinasc.microdados` as dados
WHERE ano = 2022
AND id_municipio_nascimento = '4106902'
AND classificacao_pre_natal IN ('1','2','3')
```

## Exemplo puxando os valores do dicionário via (SQL)

```
SELECT
    dict_raca_cor.valor as raca_cor,
    idade_mae,
    dict_escolaridade_mae.valor as escolaridade_mae,
    dict_estado_civil_mae.valor as estado_civil_mae
FROM `basedosdados.br_ms_sinasc.microdados` as dados
LEFT JOIN `basedosdados.br_ms_sinasc.dicionario` as dict_raca_cor
    ON dict_raca_cor.chave = dados.raca_cor
LEFT JOIN `basedosdados.br_ms_sinasc.dicionario` as dict_escolaridade_mae
    ON dict_escolaridade_mae.chave = dados.escolaridade_mae
LEFT JOIN `basedosdados.br_ms_sinasc.dicionario` as dict_estado_civil_mae
    ON dict_estado_civil_mae.chave = dados.estado_civil_mae
WHERE ano = 2022
    AND id_municipio_nascimento = '4106902'
    AND classificacao_pre_natal IN ('1','2','3')
    AND dict_raca_cor.coluna = 'raca_cor'
    AND dict_escolaridade_mae.coluna = 'escolaridade_mae'
    AND dict_estado_civil_mae.coluna = 'estado_civil_mae'
```