A Base dos Dados+: acesso fácil a dados públicos de qualidade

Rodrigo Dornelles , Matheus Valentim , Fernanda Scovino , Pedro Cavalcante

Abstract Apresentamos o projeto "Base dos Dados Mais" da ONG Base dos Dados e seu novo pacote R, recentemente disponibilizado no CRAN. Esperamos poder apresentar para a comunidade R latino-americana novas possibilidades de análise de grandes bases de dados públicos, capazes de potencializar o uso de dados abertos. Nossa plataforma, que se utiliza do motor do Google BigQuery, reúne bases de interesse pública já catalogadas, limpas, organizadas e compatíveis entre si. Pretendemos demonstrar aplicações do projeto com dados brasileiros e regionais e discutir formas de enriquecer e ampliar nosso impacto.

Palabras clave: dados públicos - dados abertos - gestão pública - políticas públicas baseadas em evidências - big data - Google BigQuery - R Stats

O que é a BD?

A Base dos Dados (BD) é uma organização sediada no Brasil que tem como missão universalizar o uso de dados, permitindo que a distância entre uma pessoa e uma análise seja apenas uma boa ideia. A BD já mapeou mais de 950 conjuntos de dados de 504 organizações.

Para atingir esse objetivo, a BD **cataloga** e **trata** bases de dados de interesse público. A catalogação consiste em gerar metadados e informações sobre uma dada fonte dos dados. Nesse processo, indica-se aos usuários onde eles podem encontrar dados de uma dada temática e informa-se o que se espera encontrar nesses dados. Tudo isso através do mecanismo de busca do site.

O tratamento de dados é o carro chefe da organização. Para essa etapa, a Base dos Dados conta com uma equipe de assistentes e voluntários que cotidianamente limpa e organiza bases de dados complexas, transformando essas bases complicadas em estruturas simples e organizadas e disponibilizando-as em seu datalake público.

Essa equipe de colaboradoras(es) realiza a limpeza e adaptação dos dados utilizando linguagens como R, Python e Stata, de acordo com o conhecimento de cada pessoa: todos os scripts são disponibilizados, juntamente com os dados originais, no repositório do projeto no GitHub a fim de assegurar a reprodutibilidade e segurança no processo de limpeza.

Dessa forma, uma base que originalmente está segmentada em vários anos, tem nomes de variáveis não inteligíveis e que talvez seja muito grande para uso direto de pessoas comuns (o que é, em regra, muito comum em se tratando de dados públicos brasileiros) se torna, após o processo, uma base fácil e amigável.

Todos esses conjuntos de dados limpos são hospedados na nuvem do Google Cloud e podem ser facilmente acessados pelo nosso pacote do R {**basedosdados**}. Essas bases já tratadas são coletivamente chamadas de **BD+**, e recebem essa tag no site, indicando que já estão limpas e prontas para uso.

Em resumo, a Base dos Dados disponibiliza gratuitamente um enorme datalake com bases já tratadas, normalizando variáveis chave entre tabelas, já em formato tidy e prontas para uso. É possível acessar dados de economia, saúde, segurança pública, orçamento, demografia, política, meio ambiente entre tantos outros, por meio do pacote {basedosdados} disponível no CRAN.

Mas por que isso é importante?

A importância disso é tornar trivial o manuseio de bases de dados complexas, porém muito úteis para a análise e elaboração de políticas públicas melhores, para a fiscalização da sociedade, jornalismo de dados e tantas outras aplicações. E, lembremos, de forma que seja possível cruzar entre si dados de eleições, economia, saúde, segurança etc já que as chaves e identificadores são padronizados em todas elas.

Embora muitas informações públicas tenham tido o acesso facilitado nos últimos anos, a leitura e manuseio delas elas ainda é complexo, instável e trabalhoso para a maior parte das pessoas interessadas.

Também, encontramos dificuldades em acessar e poder manusear bases de dados extremamente volumosas, com dezenas de GB de espaço em disco, como os dados do Censo Demográfico, a Pesquisa Nacional por Amostra de Domicílios (PNAD), Relação Anual de Informações Sociais (RAIS), microdados de vacinação contra COVID-19, Censo Escolar e outras. Esses dados, embora em tese acessíveis, são impraticáveis: além da dificuldade em interpretar e organizar os dados, há exigência de alta capacidade computacional, indisponível para a maior parte dos usuários.

Assim, através da BD+, é possível acessar facilmente esses dados, economizando muitas horas de trabalho que precisariam ser dedicadas a busca, abertura e limpeza desses dados.

O pacote {basedosdados}

Disponível no CRAN desde abril de 2021, o pacote concede acesso direto à API do BigQuery (mediada pelo {bigrquery}), permitindo sejam realizadas operações através R diretamente no serviço do Google, recebendo o resultado das operações localmente: Por exemplo:

```
# definir o billing_id correspondente ao projeto do BigQuery
basedosdados::set_billing_id("rfdornelles-bq")

# Download direto ------------------------------------------------------------

basedosdados::download(query = "
         SELECT * FROM `basedosdados.br_ana_atlas_esgotos.municipio`
         WHERE sigla_uf = 'AC'
         ",
         path =  "esgotos_exemplo.csv")


# Executar query SQL ---------------------------------------------------------

query <- "
  SELECT * FROM `basedosdados.br_ana_atlas_esgotos.municipio`
  WHERE sigla_uf = 'AC'
  "

esgostos_acre <- basedosdados::read_sql(query)
```

Conciliando todo o poder do R com o do BigQuery, o pacote a partir de sua versão 0.1.0 publicada em setembro/2021 faz uso do pacote {dbplyr} do {tidyverse}, de modo a permitir que as bases remotas sejam manuseadas mesmo sem qualquer conhecimento de SQL por parte dos usuários.

Basta criar uma variável usando uma das funções do pacote, que conectará com a base de dados remota, disponível no BiQuery, e que poderá ser manipulada usando as funções select, filter, mutate, summary entre outras e posteriormente baixadas para a memória ou para o disco.

```
# definir a tabela que vou operar:
nome_tabela <- "br_mc_auxilio_emergencial.microdados"

# fazer a conexão remota
base_remota <- basedosdados::bdplyr(nome_tabela)

# valor médio do beneficio pago por estado
# as operações são executadas apenas com comandos do tidyverse
tabela_auxilio <- base_remota %>%
  select(mes, sigla_uf, valor_beneficio, enquadramento) %>%
  group_by(sigla_uf) %>%
  summarise(
    valor_total = sum(valor_beneficio, na.rm = TRUE),
    qnt_beneficiarios = n()
  )

# coletar os dados após concluída a análise
tabela_auxilio_coletada <- basedosdados::bd_collect(tabela_auxilio)

# ou salvar em disco
basedosdados::bd_write_rds(tabela_auxilio, path = "dados_auxilio.rds", overwrite = TRUE)
```

Neste exemplo, utilizando os verbos tradicionais do {tidyverse}, foi possível em pouco menos de 1 minuto realizar uma análise na base do Auxílio Emergencial pago no Brasil, contendo literalmente milhões de linhas.

Assim, mesmo iniciantes podem acessar, filtrar, cruzar, pivotar e organizar os dados utilizando a sintaxe comum do R e com as funções largamente conhecidas do {dplyr}. Ao final, os dados poderão ser coletados já no formato desejado para elaboração de gráficos, modelos, aplicativos shiny e o que a imaginação permitir.

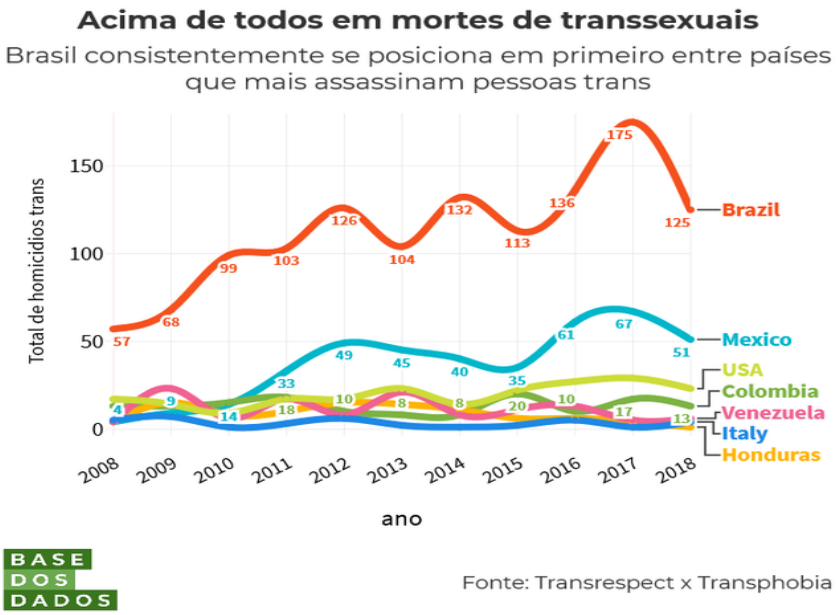
Conclusão: Algumas aplicações

Pessoas da academia, jornalistas, ativistas, pessoas da gestão pública etc podem se dedicar ao que mais importa: a análise, já que todo o trabalho de curadoria, atualização e limpeza já foi realizado por nossa comunidade.

Para que se tenha ideia do poder de tudo isso, agora é possível analisar uma base de dados de 250GB como a da RAIS (que contem dados trabalhistas e de atividade econômica de todo o país) em poucos segundos, utilizando os comandos tradicionais do {dplyr} e podendo, por exemplo, relacioná-la com dados eleitorais disponíveis na BD+ ou mesmo dados da pandemia de COVID disponíveis em bases públicas do Google.

Em breve, será possível também buscar por tabelas e metadados por meio do próprio pacote, tudo isso sem necessário sair em nenhum momento do R.

Outros vários exemplos de análises que já foram realizados pela equipe com os dados da BD+ estão disponíveis no repositório do GitHub.



Abaixo, algumas análises recentemente realizadas:

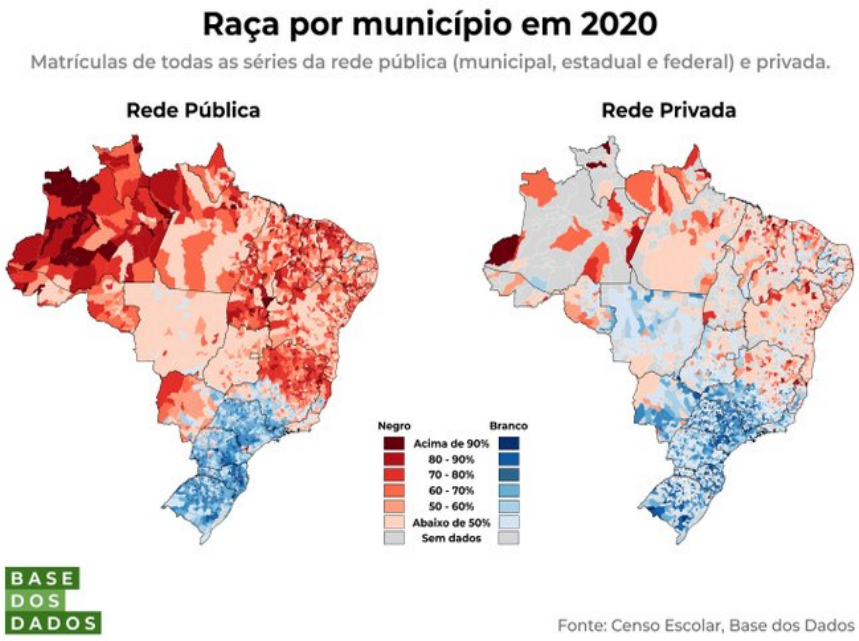


Figura 1: Gráfico comparando raça de estudantes das redes de educação pública e privada

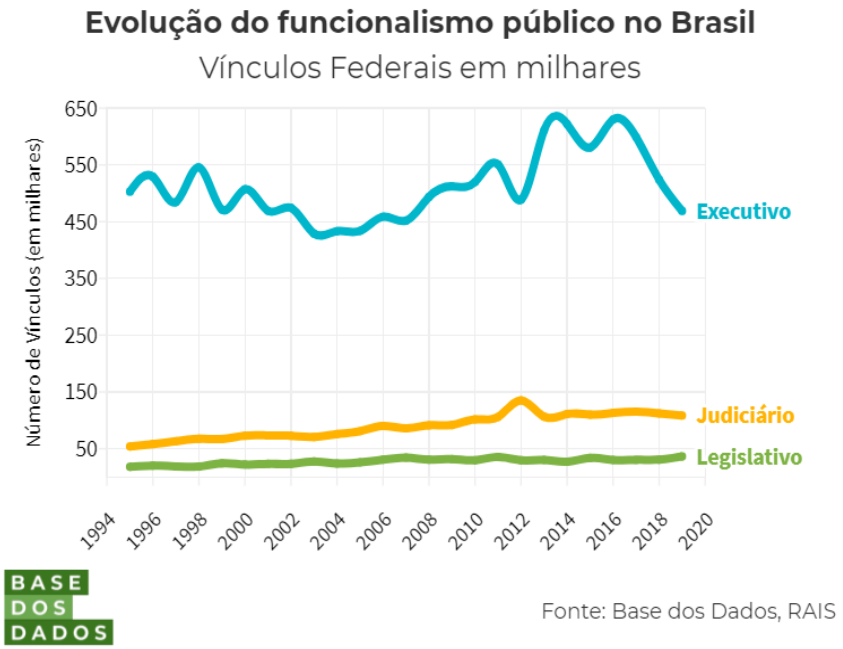


Figura 2: Quantidade de funcionários(as) públicos(as) federais vinculados aos três poderes ao longo dos últimos anos