Levantamento preliminar das fontes de dados brutos usadas pelos pesquisadores da FJP

Michel Alves - michel.alves@fjp.mg.gov.br

12/08/2021

Esse documento apresenta um levantamento preliminar e incompleto sobre as diferentes fontes e formas dos dados brutos utilizados pelos pesquisadores da FJP. O objetivo é o de dar insumos para as discussões iniciais sobre o modelo da base de dados unificada da FJP, dentro do Programa de Transformação Digital.

SIDRA IBGE

SIDRA é a sigla para Sistema IBGE de Recuperação Automática. No SIDRA, a consulta pode ser feita de três formas:

- Manualmente, diretamente no site;
- Por meio de uma URL, como por exemplo; ou
- Por meio da função `get_sidra()` do pacote `sidrar`.

Para cada tema, há uma série de tabelas com diferentes informações. Para exemplificar, considere a Pesquisa Anual do Comércio - PAC. Cada tabela é identificada por um número. A título de exemplificação, a tabela 1399 tem o título: *Número de empresas comerciais, Unidades locais com receita de revenda, Pessoal ocupado, Receita operacional líquida, Gastos com salários, retiradas e outras remunerações e Valor adicionado bruto a preços básicos segundo a divisão de comércio e faixas de pessoal ocupado.* Um exemplo de consulta é mostrado a seguir.

```
library("data.table")

entrada  <- "https://sidra.ibge.gov.br/geratabela?format=us.csv&name=tabela1399.csv&terr=N&rank=-&query=

#' Realiza a importação da tabela
tab_1399 <- fread(entrada,
                  integer64 = "numeric",
                  na.strings = c('"-"','"X"'),
                  colClasses = c(list("factor" = c(1:5))),
                  encoding = "UTF-8")
```

Consultando as informações sobre a tabela, pode-se ver que uma consulta a essa tabela retorna um total de 6 colunas, sendo elas:

1. Nivel territorial;
2. Ano;
3. Variável;

4. Divisão de Comércio;
5. Faixas de pessoal ocupado;
6. Valor.

A tabela 1407 - *Dados gerais das empresas comerciais por grandes regiões e unidades da federação de atuação das empresas e divisão de comércio e grupo de atividade* - também contém um total de 6 colunas:

1. Nivel territorial;
2. Ano;
3. Variável;
4. Grandes regiões e unidades da federação de atuação das empresas;
5. Divisão de comércio e grupo de atividade;
6. Valor.

```
entrada  <- "https://sidra.ibge.gov.br/geratabela?format=us.csv&name=tabela1407.csv&terr=N&rank=-&query=


#' Realiza a importação da tabela
tab_1407 <- fread(entrada,
                  integer64 = "numeric",
                  na.strings = c('"-"','"X"'),
                  colClasses = c(list("factor" = c(1:5))),
                  encoding = "UTF-8")
```

No entanto, a tabela 1400 - *Demonstrativo de receita das empresas comerciais por divisão de comércio, grupo e classe de atividade (CNAE 2.0)* - contém apenas 5 colunas:

1. Nivel territorial;
2. Ano;
3. Variável;
4. Divisão de comércio, grupo e classe de atividade;
5. Valor.

```
entrada  <- "https://sidra.ibge.gov.br/geratabela?format=us.csv&name=tabela1400.csv&terr=N&rank=-&query=


#' Realiza a importação da tabela
tab_1400 <- fread(entrada,
                  integer64 = "numeric",
                  na.strings = c('"-"','"X"'),
                  colClasses = c(list("factor" = c(1:5))),
                  encoding = "UTF-8")
```

Outro exemplo de tabela do SIDRA é a de número 1092 - *Número de informantes, Quantidade e Peso total das carcaças dos bovinos abatidos, no mês e no trimestre, por tipo de rebanho e tipo de inspeção* - que contém 7 colunas:

1. Nivel territorial;
2. Trimestre;
3. Referência temporal;
4. Variável;
5. Tipo de rebanho bovino;

6. Tipo de Inspeção;
7. Valor.

A tabela 3419 - *Índices de volume e de receita nominal de vendas no comércio varejista ampliado, por tipos de índice e atividades (2014 = 100)* - contém colunas também diferentes:

1. Nivel territorial;
2. Mês;
3. Variável;
4. Tipo de índice;
5. Atividades;
6. Valor.

A tabela 1618 - *Área plantada, área colhida e produção, por ano da safra e produto das lavouras* - contém as seguintes colunas:

1. Nivel territorial;
2. Mês;
3. Ano da safra;
4. Variável;
5. Produto das lavouras;
6. Valor.

PNADC IBGE

A PNADC é a Pesquisa Nacional por Amostra de Domicílios Contínua e visa acompanhar as flutuações trimestrais e a evolução, no curto, médio e longo prazos, da força de trabalho, e outras informações necessárias para o estudo do desenvolvimento socioeconômico do País. Tem diferentes periodicidades de divulgação de seus indicadores, que pode ser mensal, trimestral ou anual

Nas tabelas do SIDRA, os dados são acessados de forma direta. Diferentemente, quando trabalha-se com dados provenientes de um survey, como é o caso da PNADC, é preciso considerar o desenho do procesos amostral que gerou os dados.

Há diferentes formas de realizar a importação de dados da PNADC. A seguir serão exemplificadas.

A primeira delas é a importação online por meio da função `get_pnadc()` do pacote `PNADcIBGE`.

```
library(PNADcIBGE)
library(srvyr)
library(survey)

pnad_df <- get_pnadc(year = 2020,
                     quarter = 1,
                     vars=c("UF", "V1023", "Capital", "VD4002", "V2007", "V2009", "VD3004",
                            "VD3005", "V2010", "V1027", "V1028"),
                     design = TRUE)

pnad_srvyr <- as_survey(pnad_df)
pnad_df2 <- pnad_srvyr %>%
  filter(UF == "Minas Gerais" & (V1023 == "Capital" | V1023 == "Resto da RM (Região
                                 Metropolitana, excluindo a capital)"))


Desocupacao_cor2<- data.frame(svytable(~V2010+VD4002, pnad_df2))
```

O objeto retornado pela função `get_pnadc()` é do tipo `Stratified 1 - level Cluster Sampling design (with replacement)` e não pode ser facilmente convertido em uma tabela. Isso é realizado por meio das funções `as_survey()` e `svytable()` como mostrado no exemplo acima.

A segunda forma de importação dos dados da PNADC é de forma offline por meio da função `PnadcTidy` do pacote de mesmo nome. Para tal, é necessário fazer o download dos dados de forma manual no site do IBGE. Dois arquivos são necessários:

- `input_PNADC_trimestral.txt` - localizado em site do IBGE > Trimestral > Microdados > Documentação > Dicionario_e_Input_20210617.zip. Dentro do arquivo zip há três arquivos, dentre eles aquele que precisamos;
- `PNADC_012020.txt` - localizado em site do IBGE > Trimestral > Microdados > 2020 > PNADC_012020.zip.

Uma vez baixados e extraidos, pode-se realizar a importação, como mostrado a seguir.

```
#Importando a PNADC com a seleção de variáveis
pnadc20201 <- PnadcTidy(inputSAS="input_PNADC_trimestral.txt",
                        arquivoPnad="PNADC_012020.txt",
                        variaveis=c( "UF", "V1023", "Capital", "VD4002", "V2007", "V2009",
                                     "VD3004", "VD3005", "V2010", "V1027", "V1028" ))

#Filtrando RMBH
pnadc20201 <- filter(pnadc20201, pnadc20201$UF == 31)
pnadc20201 <- filter(pnadc20201, pnadc20201$V1023 == 1 | pnadc20201$V1023 == 2)

#Declarando as variáveis categóricas

pnadc20201$VD4002 <- factor(pnadc20201$VD4002, label=c("Pessoas ocupadas", "Pessoas desocupadas"),
                            levels=1:2)
pnadc20201$V2010 <- factor(pnadc20201$V2010, label=c("Branca", "Preta", "Amarela", "Parda",
                                                     "Indígena", "Ignorado"), lev = c(1,2,3,4,5,9))


pnadc202012 <-
  svydesign(
    ids = ~ UPA ,
    strata = ~ Estrato ,
    weights = ~ V1028 ,
    data = pnadc20201 ,
    nest = TRUE)

#Taxa de desocupação por cor ou raça
Desocupacao_cor<- data.frame(svytable(~V2010+VD4002, pnadc202012))
```

O exemplo acima gera a mesma tabela gerada no exemplo de importação online. O objeto retornado pela função `PnadcTidy()` é uma tabela com 15 colunas. No entanto, ainda é necessário aplicar os pesos do desenho amostral, o que é feito pela função `svydesign()`. O objeto retornado por essa função é do tipo `Stratified 1 - level Cluster Sampling design (with replacement)`.

ANAC

Os dados estatísticos da ANAC - Agência Nacional de Aviação Civil - estão disponíveis no site. São apresentados em diferentes formas, como arquivos com extensão xlsx, csv ou pdf.

A título de exemplificação, considere os dados obtidos por meio do site da ANAC > Assuntos > Dados e Estatísticas > Mercado do Transporte Áereo > Base de Dados Estatísticos do Transporte Aéreo > Dados Estatísticos do transporte aéreo do Brasil.

O arquivo baixado, (depois de extraído) corresponde a um arquivo de nome 'resumo_anual_2020" com extensão csv com 38 colunas e mais de 26 mil linhas. Algumas dessas colunas são mostradas a seguir.

- Empresa
- Ano
- Mês
- Aeroporto de Origem - SIGLA
- Aeroporto de Origem - NOME
- Aeroporto de Destino
- Natureza
- Grupo de Vôo
- Passageiros Pagos
- Passageiros Grátis
- Distância Voada
- Bagagem (Kg)

Apesar da grande quantidade de informações disponibilizadas no arquivo, normalmente só algumas das colunas e linhas são utilizadas.

SEDESE - MG

Dentro os dados divulgados pela SEDESE - Secretaria de Estado de Desenvolvimento Social - está o Relatório dos Indicadores Preliminares do ICMS - Critério Esportes, disponível aqui. Esse relatório, que é um arquivo com extensão pdf, contém uma tabela com a *Listagem dos Municípios pontuantes no índice de esportes do ICMS solidário*, referente ao ano de 2019, com 4 colunas:

- Ranking;
- Município;
- Nota final;
- Percentual.

CAGED

A importação dos dados do CAGED - Cadastro Geral de Empregados e Desempregados - pode ser realizada de forma offline, acessando o esse link (o acesso só é realizado quando se utiliza o navegador Internet Explorer). Os dados utilizados nesse exemplo estão disponíveis no Diretório FTP > Novo CAGED > Estabelecimentos > 2020 > Dezembro > CAGEDESTAB202001.7z. Uma vez os dados baixados e descompactados, pode-se realizar a importação dos dados como mostrado a seguir.

```
cagedjan2020<- read.table("CAGEDESTAB202001.txt", head=T, sep=";", encoding = "UTF-8")
```

O objeto gerado é uma tabela de 13 colunas e mais de 830 mil linhas. As colunas são apresentadas a seguir.

- Competência;
- Região;
- UF;
- Município;

- Seção;
- Subclasse;
- Admitidos;
- Desligados;
- Fonte desligamento;
- Saldo movimentação;
- Tipo empregador;
- Tipo estabelecimento;
- tamestabjan.

Normalmente faz-se uma seleção de colunas. Com relação às observações (linhas), é comum se trabalhar com todas elas. No entanto, usa-se as linhas para se obter, como por exemplo, o saldo de empregos por unidade federativa para um determinado mês.

ANP

Os dados da ANP - Agência Nacional de Petróleo - podem ser importados de maneira offline. Para tal, deve-se acessar o site da ANP. Para este exemplo, será acessado os dados disponíveis em Site da ANP > Vendas de derivados de petróleo e biocombustíveis > Vendas Etanol Hidratado (metros cúbicos) 1990-2020. O arquivo baixado tem a extensão csv e possui 18 colunas e mais de 800 linhas.