



International Seminar on Statistics with R

R, Python e Dados abertos no telegram: redes de colaboração por mais controle social

Fernando Almeida Barbalho - Secretaria do Tesouro Nacional



Fórum Romano - Roma 2017

Título:	Emergência de um campo de ação estratégica : o caso de política pública sobre dados abertos
Autor(es):	Barbalho, Fernando Almeida
Orientador(es):	Medeiros, Janann Joslin
Assunto:	Políticas públicas - avaliação Dados abertos Transparência na administração pública Governo eletrônico - avaliação



# LA THEORIE DES CHAMPS DE PIERRE BOURDIEU

a theory  
of fields

NEIL FLIGSTEIN | DOUG McADAM







Oceanário - Lisboa 2019

# As funções da Transparência

Controle Social

Confiança

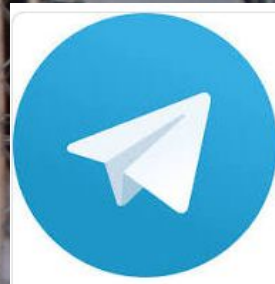
Desenvolvimento de produtos e  
serviços







Festival de Montreux - 2011



Por que misturar?







Schloss Platz - Berlin 2007

Os capitais como recursos para as arenas de disputa e colaboração:

- Capital Econômico
- Capital Social
- Capital Simbólico
- Capital Técnico



NYC 2015

Capital Econômico

Free software and governments | OCS-Mag  
ocsmag.com



Free software and governments | OCS-Mag  
ocsmag.com



## Capital Econômico

### Pt-BR Data Science & Python

Ou prontamente, aprender, sendo auto-didata, aí tem uma infinidade de conteúdo, cursos, e material. Após aprender, montar um portfólio e botar a cara no mercado. Talvez alguém experiente possa lhe ajudar, lhe dando uma direção. Um coach ou uma pessoa de confiança que já trabalha na área.

# Capital Econômico

## Negócios com Dados Abertos

**Fernando Barbalho**

Muito bom o trabalho Lucas. Dentro da proposta aqui do grupo ...

Obrigado Fernando! Quando criei essa análise meu objetivo não era monetizar, mas sim mostrar a população o poder dos dados abertos! Atualmente estou utilizando essa análise para enriquecer o meu portfólio, o que tem me rendido bons contatos! Claro, podemos sim desenvolver uma narrativa mais elaborada e publicar em uma plataforma como o Medium, assim como também no Kaggle. Hmmm, esse modelo via google ads não sei se iria funcionar, precisaríamos de muito tráfego, pra validar poderíamos fazer um MVP.

# Capital Econômico

**R Brasil**

Boa tarde pessoal.

Sou engenheiro por formação mas estou estudando para entrar profissionalmente na área de engenharia de dados.

A maioria desses profissionais tem graduação na área de TI. Devido a isso gostaria de saber como que é a absorção de profissionais de outras áreas. Vale a pena fazer uma graduação na área ou certificações como a da DSA, Udemy, Udacity e etc tem o mesmo peso no currículo? Há a necessidade de um portfólio? Grato a todos.





Brasília: Congresso Nacional - 2016

Capital Social



**R Brasil**

1031 membros



**Pt-BR Data Science & Python**

3939 membros



**PyData BSB**

810 membros



**Dados Abertos .BR**

1440 membros



## Capital Social



THE UNIVERSITY OF CHICAGO PRESS JOURNALS

[American Journal of Sociology](#) / [Vol. 78, No. 6](#) / [The Strength of...](#)



JOURNAL ARTICLE

### **The Strength of Weak Ties**

Mark S. Granovetter

*American Journal of Sociology*

Vol. 78, No. 6 (May, 1973), pp. 1360-1380

Published by: [The University of Chicago Press](#)

<https://www.jstor.org/stable/2776392>


Page Count: 21


**Topics:** [Sociometrics](#), [Friendship](#), [Social networks](#), [Communities](#), [Innovation adoption](#), [Social structures](#), [Rumors](#), [Social theories](#), [Community organizing](#), [Biophysics](#)

[Give feedback](#)




# Capital Social

**Neto Ferraz**  
last seen at 11:19 AM





**Info**


@nferraz  
Username


Notifications  
On 


**Shared content**


 Links 9


 Voice messages 2

 Groups in common 8


 **Groups in common**


**Administradores - DA Br**  
private group


**Open Data Day Brasília**  
private group


**NA**


**Negócios com Dados Abertos**  
public group

**PyData BSB**  
public group

**Pt-BR Data Science & Python**  
public group

**R Brasil**  
public group

**Software Livre no Setor Público**  
public group

**Dados Abertos .BR**  
public group



7<sup>e</sup> Arr.  
PLACE  
DOR ALLENDE  
1908 - 1973  
DE LA RÉPUBLIQUE DU CHILI

2010



Capital simbólico

Confiança no Estado

Legitimidade do  
Estado





# Capital simbólico



**dados.gov.br**

PORTAL BRASILEIRO DE DADOS ABERTOS



**BANCO CENTRAL  
DO BRASIL**

Portal de Dados Abertos



**TESOURO NACIONAL**  
TRANSPARENTE

## Portal da Transparência

CONTROLADORIA-GERAL DA UNIÃO



Chichén Itzá- México: 2014



# Capital técnico

"Your fingers are just like... they type R code without you thinking about it. It just flows out of your fingers. [...] It's just completely subconscious. R flows out of your fingers and everything just works."

~ HADLEY WICKHAM, may his beard grow ever longer ~

The bad  
a long  
frustr  
som  
tempo  
from kn  
some  
without

"I think there's nothing particularly special about R, other than it just being the greatest software on Earth."

~ AMANDA COX, Data Editor of The New York Times ~



Black Hole Researcher Katie Bouman  
vice.com





# Capital técnico

Beatriz Milz é co-organizadora do R-Ladies São Paulo, bacharel em Gestão Ambiental (EACH/USP), mestre em Ciências (UNIFESP) e atualmente doutoranda em Ciência Ambiental (IEE/USP). É pesquisadora no Projeto Temático FAPESP "Governança Ambiental na Macrometrópole Paulista face à variabilidade climática" (Processo: 2015/03804-9), e Assessora Editorial da Revista Ambiente & Sociedade.

Haydee Svab é co-fundadora dos grupos RLadies-São Paulo, do PoliGNU (Grupo de Estudos de Software Livre da Poli-USP) e da PoliGen (Grupo de Estudos de Gênero da Poli-USP). É Engenheira Civil, mestra em Engenharia e Planejamento de Transportes (Poli-USP), especialista em Democracia Participativa, Repúblicas e Movimentos Sociais (UFMG) e doutoranda em Smart Cities pelo IME-USP. Atualmente é consultora do Banco Mundial, CEO da ASK-AR (consultoria em análise de dados), membro do Conselho Deliberativo da AEAMESP (Associação dos Engenheiros e Arquitetos de Metrô) e da comunidade Transparência Hacker.

## Julio Trecenti

### President at CONRE-3a Região

Considera-se um Faxineiro de Dados. Sócio-fundador da Curso-R. Doutorando em Estatística pelo IME-USP. Secretário-geral da Associação Brasileira de Jurimetria (ABJ). Sócio-fundador da Terranova Consultoria. Trabalha com web scraping, arrumação de dados, construção de modelos preditivos, APIs, pacotes em R e dashboards em Shiny. Coordenador e ministrante de diversos cursos sobre R, ciência de dados e jurimetria.

Bea Milz

Haydee Svab

Bruna Wundervald

Julio Trecenti



R Brasil

1031 membros



Pt-BR Data Science & Python

3959 membros



Dados Abertos .BR

1443 membros

Bruna Wundervald

Ph.D. Candidate in Bayesian Machine Learning & R-Ladies Co-organiser

Bruna is a Ph.D. candidate in Bayesian Machine Learning at the Hamilton Institute, in the National University of Ireland Maynooth. Member and co-organiser of the Curitiba and São Paulo R-Ladies chapters in Brazil, also involved with the worldwide community. Founder and developer of the R-Music organization, that promotes the study of music information retrieval in R. Previously, she obtained her BSc in Statistics at the Paraná Federal University, where she worked in a diversity of extension projects involving R and statistics. Interested in ML in general, package and dashboards building, text mining and APIs. Her work is now especially focused on the development of new methods for Bayesian machine learning using both R and python, as well as multivariate statistics, variational inference, and MIR.



Arquipélago de Anavilhanas – Floresta Amazônica: 2013

## Caso 1: Tesouro Nacional

R Brasil 1031 membros

Mostrar mensagens rece



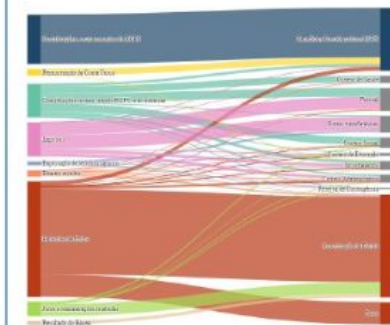
Charles Lula da Silva admin

@FBarbalho, vc já viu isso?

<https://medium.com/@fernandobarbalho/suporte-do-rstats-às-iniciativas-de-transparência-do-tesouro-nacional-brasileiro-dfdd5e1ab831> 🙏🙏🙏

### Suporte do rstats às iniciativas de transparência do Tesouro Nacional brasileiro

No final de 2018 a Secretaria do Tesouro Nacional brasileiro disponibilizou a nova versão do portal Tesouro Transparente. No portal estão...



Dados Abertos .BR 1443 membros

Mostrar mensagens rece

Dá pra contornar o captcha, n?



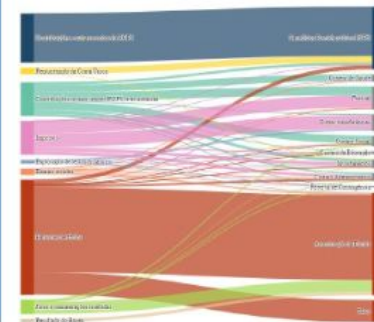
Fernando Barbalho admin

Artigo que mostra o trabalho da equipe de Ciência de Dados na Secretaria do Tesouro Nacional. O foco do artigo é nos produtos desenvolvidos utilizando R.

<https://link.medium.com/g83yyy7oBU>

### Suporte do rstats às iniciativas de transparência do Tesouro Nacional brasileiro

No final de 2018 a Secretaria do Tesouro Nacional brasileiro disponibilizou a nova versão do portal Tesouro Transparente. No portal estão...



<https://link.medium.com/g83yyy7oBU>



## Caso 1: Tesouro Nacional

<b>VIEWS BY TRAFFIC SOURCE</b>	<b>3.7K</b>
<b>Medium</b> ?	<b>9%</b>
<hr/>	
<b>External referrals</b>	<b>91%</b>
email, IM, and direct	1.6K
<a href="#">facebook.com</a>	984
<a href="#">twitter.com</a>	330
rsci.app.link	176
datascienceacademy.com.br/social	49
google.com	39
tesouro.sharepoint.com	34
<a href="#">linkedin.com</a>	24
Android app	20
embrapa.br	14
All other external referrals	52

## Caso 1: Tesouro Nacional

<https://medium.com/tchiluanda/levando-o-patinho-feio-ao-nirvana-uma-hist%C3%B3ria-de-amor-e-%C3%B3dio-entre-pdf-e-dados-abertos-37f9b62ec0b0>

Dados Abertos .BR 1443 membros

Mostrar mensagens rece



**Fernando Barbalho** admin

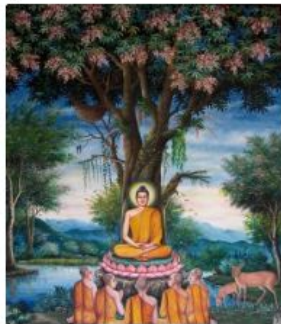
edita

Caros, lembram da discussão sobre pdf que rolou aqui há algumas semanas? estava sem poder participar no momento, mas agora escrevi um texto no medium que relata nossa experiência de como utilizamos o pdf no contexto de dados abertos. Segue o link:

<https://medium.com/@fernandobarbalho/levando-o-patinho-feio-ao-nirvana-uma-história-de-amor-e-ódio-entre-pdf-e-dados-abertos-37f9b62ec0b0>

### Levando o patinho feio ao nirvana: uma história de amor e ódio entre pdf e dados abertos

Não há como negar, o pdf é o patinho feio do campo de dados abertos. No modelo 5 estrelas de TBL ele é até considerado, mas aparece como...



Lucas

07:31:01

Fernando Barbalho

"Levando o patinho feio ao nirvana: uma história de amor e ódi...

Muito interessante. Mas que distância para a maior parte dos órgãos do governo.

De qualquer forma, parabéns pelo trabalho

Fernando Barbalho

07:31:55

Lucas

Muito interessante. Mas que distância para a maior parte dos ór...

Obrigado Lucas

Neto Ferraz

07:57:12

Sou fã da equipe do GT da STN @FBarbalho @tiagombp ! Não sei se os demais estão aqui. Mas estendo a minha admiração aos demais integrantes da equipe!

Sillas admin

08:38:06

Lucas

Muito interessante. Mas que distância para a maior parte dos ór...

Como o serviço público seria melhor se em todas as estatais houvesse equipes tão preocupadas em fazer um bom trabalho como a do @FBarbalho. O trabalho de vcs é fantástico!

Rodrigo Almeida

08:40:48

👏👏👏👏

Bea Milz

09:23:58

Lucas

Muito interessante. Mas que distância para a maior parte dos ór...

Também pensei isso! Achei super legal o que fizeram com shiny. Mas é distante da realidade de outros órgãos públicos

E nesse caso eram relatórios, e quando os órgãos disponibilizam as tabelas em pdf ? Complicado

Ainda to passando uns perrengues tentando abrir alguns às vezes

Mas novamente, parabéns pelo trabalho de vocês

Espero que vire tendência eventualmente ahshshhs

Ariel Levy

10:01:42

Fernando Barbalho

"Levando o patinho feio ao nirvana: uma história de amor e ódi...

Muito bom



Capital simbólico e social

**Fernando Barbalho**

@nferraz e @sillastg a equipe agradece às mensagens carinhosas de vocês. Como na história de Quincas Borba que recupera toda a série de contribuições no processo de levar o frango à mesa, temos um grande e impagável débito a várias pessoas e equipes, mas principalmente a este grupo R-Brasil onde aprendemos muito (para não dizer quase tudo) do que produzimos atualmente.



Vista noturna Rio de Janeiro: 2012



## Caso 2: Convênios do Governo Federal

### Análise exploratória dos dados de convênios do governo federal (Parte 1)

Os dados utilizados nesta análise foram obtidos no Portal da Transparência, na [área de dados abertos](#), e no momento do download estavam atualizados até 19 de abril de 2019.

#### Tópicos

1. **Importando os dados**
2. **Entendendo a estrutura dos dados**
  - Qual é o tamanho da base?
  - Quais são as colunas e que tipo de dados contêm?
  - Que tipos de convênios?
  - Que status de convênios são possíveis?
3. **Tratamento dos dados**
  - Convertendo os formatos de data
  - Removendo os valores discrepantes
4. **Cruzando com a base de ordem bancária**
5. **Algumas análises básicas**
  - Quanto foi liberado por tipo de convênio em toda a série histórica?
  - Quanto foi liberado por ano no total?
6. **Algumas sugestões de melhoria para publicação dos dados**

Os dados utilizados nesta análise foram obtidos no Portal da Transparência, na [área de dados abertos](#), e no momento do download estavam atualizados até 19 de abril de 2019.

[https://github.com/campagnucci/dados-agentes-gov-br/blob/master/analise\\_convencios.ipynb](https://github.com/campagnucci/dados-agentes-gov-br/blob/master/analise_convencios.ipynb)

## Caso 2: Convênios do Governo Federal



**Fernanda Campagnucci**

editado

Olá pessoal! Depois da discussão sobre a atualização dos dados do Portal da Transparência do governo federal, comecei a fazer algumas análises simples sobre o conteúdo e forma de publicação. A primeira, exploratória, foi sobre os dados de convênios. O maior ponto de atenção é a ausência de uma coluna importante, que é o tipo de conveniente (osc, ente público, pessoa física, etc). Acredito que tenha sido um lapso. Também tem problemas mais gerais de qualidade do dado (80% sem info sobre o tipo de instrumento), faltaria aí uma contextualização sobre essa variável. Vou avançar em outras análises quando der tempo.. sugestões são bem vindas (e já encaminhei pro email da CGU).

<https://github.com/campagnucci/dados-abertos-gov-br>

GitHub

**[campagnucci/dados-abertos-gov-br](https://github.com/campagnucci/dados-abertos-gov-br)**

Tratamento e análise de dados abertos do Governo Federal do Brasil -  
[campagnucci/dados-abertos-gov-br](https://github.com/campagnucci/dados-abertos-gov-br)



Capital social



**Dados Abertos .BR**

1443 membros

BM

**Bruno Morassutti** admin

15:35:44

**Fernanda Campagnucci**

Olá pessoal! Depois da discussão sobre a atualização dos dados ...  
Interessante, Fernanda. Fiz uma inclusive um questionamento via eSIC sobre as formas pelas quais o cidadão pode dar feedbacks ou participar na definição de prioridades do Portal. Assim que tiver um retorno, compartilharei

O ideal mesmo seria de a CGU passasse a adotar meios mais abertos para o desenvolvimento de plataformas em geral. Para uma entidade que trabalha com transparência, eles infelizmente ainda são muito fechados

15:36:55

CH C convidou **Fabiano Nascimento**



**Nitai Bezerra da Silva** admin

15:46:31

**Fernanda Campagnucci**

Seria legal ter um canal de feedback sobre os dados no portal. ...

Fernanda, nos últimos tempos eu venho vislumbrando que as ouvidorias têm servido cada vez mais como um canal de feedback sobre qualquer coisa. Elas são bem empoderadas na maioria dos casos que conheci. E com as integrações que a OGU/CGU vem implementando tenho visto elas no centro desse processo. O ideal é que cada plataforma criasse seus fóruns de diálogo com suas bases de usuários. Enquanto isso não acontece acho que deveríamos explorar os canais com as ouvidorias



**Fernanda Campagnucci**

15:52:21

**Bruno Morassutti**

O ideal mesmo seria de a CGU passasse a adotar meios mais ab...

Uma coisa que essa história me fez pensar é sobre a necessidade de abriremos também os scripts ou procedimentos usados na elaboração dos gráficos dos portais de transparência. Com exceção de alguns anos, eu simplesmente não consigo reproduzir o gráfico de valores calibrados e liberados dos convênios com os dados



## Caso 3: Mobilidade urbana em Maceió

Aeroporto  
Internacional  
Zumbi dos  
Palmares

AL-210 Rio Largo

AL-105

<https://medium.com/pizzadedados/utilizando-dados-abertos-e-ci%C3%A2ncia-de-dados-na-mobilidade-urbana-371a4c591639>

# Utilizando Dados Abertos e Ciência de dados na mobilidade urbana

Entendendo a cidade de Maceió através de imagens



Armando Barbosa  
Apr 16 · 4 min read

Santa Luzia  
do Norte

424

AL-401  
Coqueiro Seco

Mundaú Lagoon

104

316

316

AL-101

Maceió

PONTA VERDE

424

AL-101

AL-101

Google

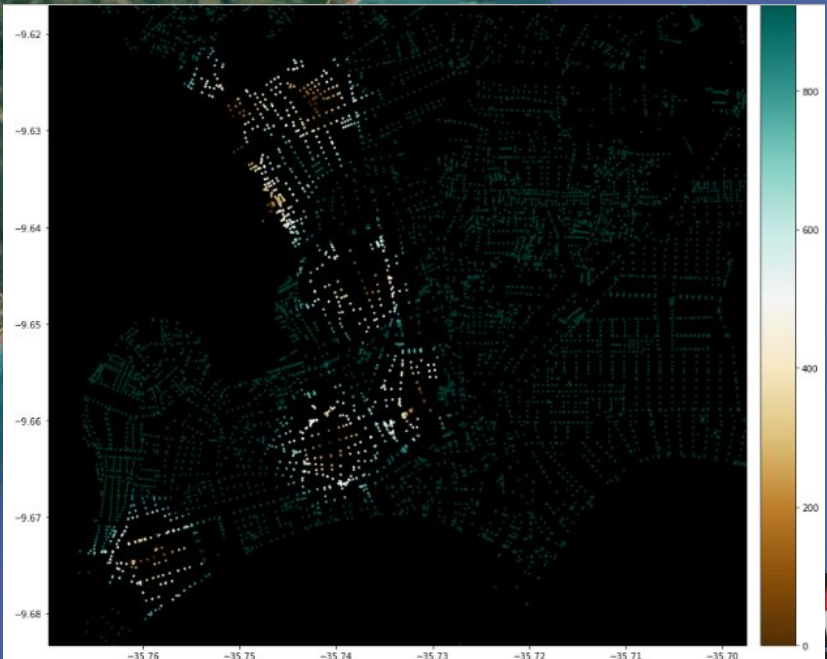


Figura 3 — Concentração de hospitais em Maceió

Na Figura 3, podemos notar uma concentração maior de hospitais do lado oeste. Consequentemente a população mais distante precisará se deslocar mais para chegar a um desses hospitais.



## Caso 3: Mobilidade urbana em Maceió

Armando Barbosa

10:07:27

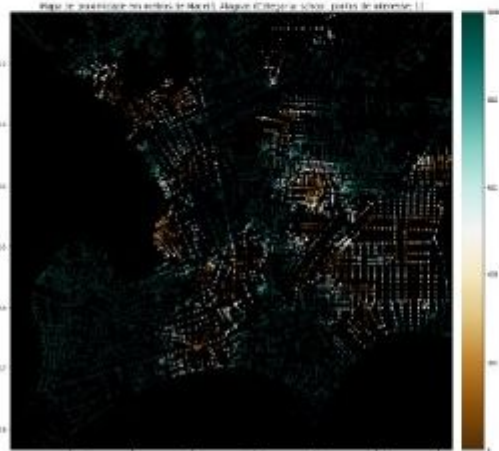
Olá pessoal,  
artigo quentinho!

Neste artigo utilizamos dados do OpenStreetMap para analisar a mobilidade urbana, corre lá e confere.

<https://medium.com/@armandobs14/utilizando-dados-abertos-e-ciencia-de-dados-na-mobilidade-urbana-371a4c591639>

### Utilizando Dados Abertos e Ciência de dados na mobilidade urbana

Entendendo a cidade de Maceió através de imagens



Augusto Batista admin

10:19:05

Armando Barbosa

Olá pessoal, artigo quentinho! Neste artigo utilizamos dados do ...  
Não conhecia essas bibliotecas Python osmnx e pandana. Pelo que vi essa última funciona junto com o pandas para usar os dados do Open Street Map. Parabéns pelo artigo e obrigado por compartilhar!

Capitais técnico e econômico



**Dados Abertos .BR**

1443 membros

NA

**Negócios com Dados Abertos**

32 membros

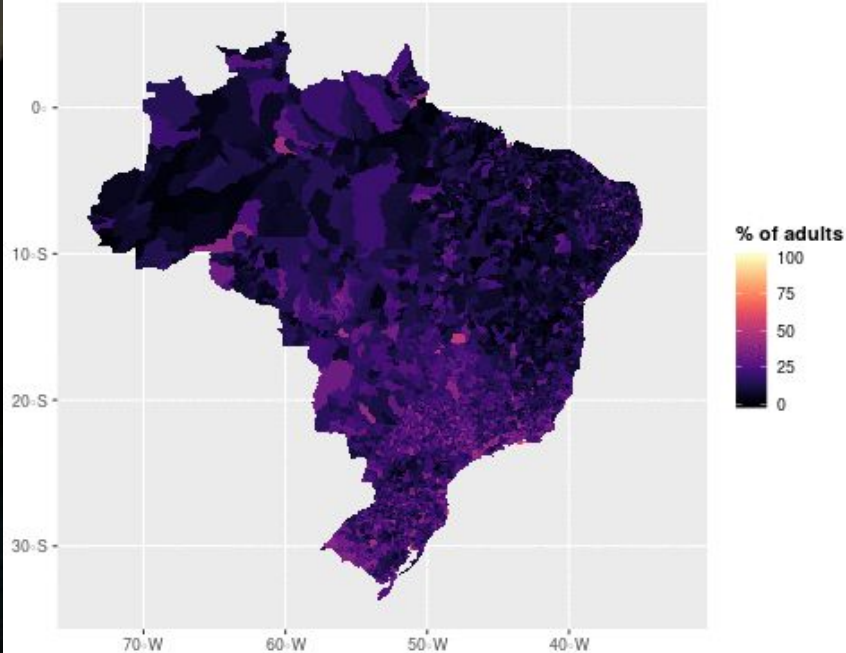
Google



Por do Sol em Fortaleza: 2016

## Caso 4: “Light of education”

% of adults(+18) with fundamental education: 1991



Source: UNPD, made available by Abj

## Insights from data visualization



Charles Novaes de Santana

Mar 11 · 2 min read

the country. On the other hand, the animation shows how this proportion is increasing along last years and how the “light of education” is spreading across the country.

<https://medium.com/@charlesnovaes/insights-from-data-visualization-edf02febfa9e>



## Caso 4: "Light of education"

Se voces repararem no código que mandei acima, tentei manter a mesma paleta usando a função `scale_fill_viridis_c(option = "plasma", values = rescale(seq(from=0.3,to=1,by=0.1)), breaks = seq(from=0.3,to=1,by=0.1))`

Agradeço por qualquer ajuda!

Sillas admin

Charles Lula da Silva

Agradeço por qualquer ajuda!

e se vc, ao invés de criar um gráfico separado filtrando o year no início, fazer um facet por year?

eu quero tentar fazer isso, mas vai demorar pra eu instalar o sf

Neto Ferraz

Charles Lula da Silva  
Foto

map-Texa... .png 189 KB  
Baixar Abrir

Seu trabalho tá ficando bem interessante. Mas confesso que prefiro a **multi-hue color scale**, como uma Viridis, por exemplo, que você estava adotando. Acredito que visualmente fica mais prontamente identificável os extremos. Além disso, algo que me veio a cabeça é quando estamos tratando de indicadores socioeconômicos algo interessante de abordar é o **gap** entre capital e interior. E para esse

Capital técnico

Sillas admin

Charles Lula da Silva

eu imagino que assim vai funcionar. Mas eu queria fazer uma a... mas o gganimate vai manter a mesma escala, não?

e eu acho o facet uma alternativa melhor ao gganimate

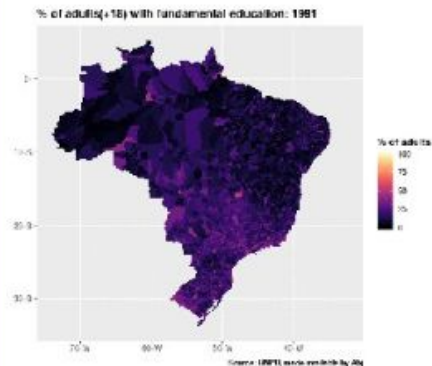


Fernando Barbalho

"Insights from data visualization" by Charles Novaes de Santana  
<https://link.medium.com/Hbf1jULRYU>

### Insights from data visualization

Data visualization is a great way of getting insights from data. This is particularly true when we consider geo-referenced data. Let me...



O Medium me sugeriu a leitura acima. Dizem que é de um cara dos bons nas ciências de dados



R Brasil

1031 membros



Centro de São Paulo: 2013



## Caso 5: CNPJ

### Pacote desenvolvido para tratar e organizar a Base de dados do Cadastro Nacional da Pessoa Jurídica (CNPJ)

#### O Projeto

O pacote `qsacnpj` é uma das ferramentas utilizadas no projeto de “Transparência das Contas Públicas”, desenvolvido e executado pelo Observatório Social do Brasil - Município de Santo Antônio de Jesus - Estado da Bahia.

Atualmente, o projeto é composto das seguintes ferramentas:

Pacotes desenvolvidos em Linguagem R para realizar Web Scraping e tratamento de dados:

<https://brasil.io/dataset/socios-brasil/socios>



### Sócios das Empresas Brasileiras

Quadros societários e de administradores das pessoas jurídicas brasileiras.

Fonte original: [Receita Federal do Brasil](#)

Libertado por: [Álvaro Justen](#)

Código-fonte: <https://github.com/turicas/socios-brasil>

Licença: [Creative Commons Attribution-ShareAlike 4.0 International \(CC BY-SA 4.0\)](#)

Links relacionados: [Dicionário de dados da qualificação dos sócios, Decreto nº 8.777, de 11 de maio de 2016](#)

<https://github.com/georgevbsantiago/qsacnpj>



## Caso 5: CNPJ

Dados Abertos .BR 1444 membros

Mostrar mensagens recentes



BM

**Bruno Morassutti** admin

**Cadu Vieira**

O MF enviou hoje para mim, pelos Correios, um pendrive com o...  
Me mandaram também o comprovante de remessa. Estou no aguardo também

Ficaram de me encaminhar a base do CNPJ completa

JG

**Josir Gomes**

**Cadu Vieira**

O MF enviou hoje para mim, pelos Correios, um pendrive com o...



GS

**George Santiago** admin

**Cadu Vieira**

O MF enviou hoje para mim, pelos Correios, um pendrive com o...  
Perplexo\*2

**Bruno Morassutti**

Ficaram de me encaminhar a base do CNPJ completa  
Sensacional!!!

Essa base de dados expande para muitas possibilidades de cruzamentos de dados.

BM

**Bruno Morassutti** admin

Pessoal, parece que a Receita colocou no ar um link direto para o arquivo que recebemos. Curiosamente, foi ao ar na sexta-feira, 23/11/2018, mesmo dia que o @caduvieira e eu recebemos os arquivos por correio.

<https://idg.receita.fazenda.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj>

Dados públicos CNPJ



**Rodrigo Brasil**

Estava lendo o PDF, e diz que o CPF dos sócios está mascarado, ocultaram os 3 primeiros e os 2 últimos dígitos

BM

**Bruno Morassutti** admin

**Rodrigo Brasil**

Estava lendo o PDF, e diz que o CPF dos sócios está mascarado,...  
Sim... essa foi a recomendação da CGU ao deferir os recursos

## Caso 5: CNPJ

R Brasil 1031 membros

Mostrar mensagens recentes

GS

**George Santiago** 14:01:26

Galera da Comunidade R. O Cadu Vieira disponibilizou a Base do CNPJ e Quadro Societário no grupo Dados Abertos .Br. O arquivo representa um desafio enorme de Tratamento de Dados (ETL)..

Link da Base de Dados:  
[http://bit.ly/CNPJ\\_QSA](http://bit.ly/CNPJ_QSA) 14:01:37

Download icon

**Documentação\_559....pdf** 94 KB 14:02:10

Baixar

O dicionário de dados 14:02:13

Será que tem como trabalhar isso no R? kkkkk 14:02:31

F.K032001K.D81106A

87.007.869.830

 14:02:43

Olha o tamanho da "criança". 14:02:49

Depois de quase um ano "brigando" pelo acesso desta Base de Dados junto a Receita Federal, usando a LAI... finalmente enviaram. 14:03:33

Contudo, o desafio agora é de ETL para organizar essa Base de Dados 14:03:46

Sillas admin 14:10:22

**George Santiago** 14:10:52

Galera da Comunidade R. O Cadu Vieira disponibilizou a Base do... Já tem pacote pra isso

[https://www.google.com/url?q=https://www.curso-r.com/blog/2018-05-13-rfbcnpj/&sa=U&ved=2ahUKEwjKxNao9-\\_eAhVLIZAKHXQhC1gQFjAAegQIARAB&usg=AOvVaw232MzkMpxrtbOmR2cnxxWV](https://www.google.com/url?q=https://www.curso-r.com/blog/2018-05-13-rfbcnpj/&sa=U&ved=2ahUKEwjKxNao9-_eAhVLIZAKHXQhC1gQFjAAegQIARAB&usg=AOvVaw232MzkMpxrtbOmR2cnxxWV) 14:10:52

**Curso-R** 14:11:10

Durante nosso último curso de introdução a programação em R, o Reinaldo Chaves me pediu ajuda para carregar os dados de CNPJs da Receita Federal. Eu m...

**George Santiago** 14:11:50

Essa é uma base de dados nova

A anterior, a Receita Federal não tinha enviado o CPF dos sócios, nem dados sobre endereço, CNAE... Já nesta tem (ou deveria ter). 14:11:50

**Sillas** 14:12:25

<https://www.google.com/url?q=https://www.curso-r.com/blog/...>

E no exemplo do pacote, os arquivos estão separados por Estado. Desta vez, a RFB enviou um único arquivo num PenDrive.



## Caso 5: CNPJ

Fernando Gomes Jr

17:17:38



**Cadu Vieira** 25 de nov de 2018 13:00:00

[http://bit.ly/CNPJ\\_QSA](http://bit.ly/CNPJ_QSA)



**Rodrigo Brasil** 25 de nov de 2018 16:16:07

No google drive está indicando 5GB



**Reinaldo Chaves** 25 de nov de 2018 16:39:25

sim, descompacte e tem 87 GB, é um fixed width. obrigado  
no PDF tem as posições das informações, certo? a partir da página 4  
por exemplo, o CNPJ começa na posição 4 e vai até 18



**Álvaro Justen** 25 de nov de 2018 17:08:46

Eu baixei aqui o que o @jedibruno subiu e estou catalogando os  
metadados dos fixed-width files para que a extração e conversão a  
partir daí seja automática

Fernando Gomes Jr

17:19:16

**George Santiago**

Contudo, o desafio agora é de ETL para organizar essa Base de ...  
O pessoal que tem interesse em trabalhar nesse dataset pode trocar  
uma ideia com o Alvaro Justen, do [brasil.io](https://brasil.io)

**Fernando Gomes Jr**

17:20:49

Eu baixei aqui o que o @jedibruno subiu e estou catalogando os...  
Ele disse que já começou, então a galera pode conversar com ele e  
evitar o retrabalho:

<https://t.me/turicas>

R Brasil 1031 membros

Mostrar mensagens recentes



**Julio Trecenti** admin

vamos esperar o turicas subir :D

depois a gente empacota

eu dei uma olhada no arquivo e é bem chato de mexer, deve  
demorar algumas horas pra entender como ele funciona e  
processar.. melhor deixar o turicas mexer e evitar retrabalho



**José De Jesus**

**Tomás Barcellos**

José, onde está? No github?

Não estou na frente do meu pc mas tá no meu gist. Qdo chegar,  
partilho.



**Fernando Gomes Jr**

**Julio Trecenti**

eu dei uma olhada no arquivo e é bem chato de mexer, deve de...

Ele avisou no grupo de dados abertos que quando terminar vai  
disponibilizar o código no GitHub e os dados em outro lugar



**Julio Trecenti** admin


yep



## Caso 5: CNPJ

**Dados Abertos .BR** 1444 membros

Mostrar mensagens recentes


**Álvaro Justen** admin19:03:59

Dados e script liberados:  
<https://twitter.com/turicas/status/1067161458214092802> cc @caduvieira

Twitter

Álvaro Justen

Recebemos o quadro de sócios e administradores das empresas brasileiras via LAI e eu já extraí. Dados em:  
<https://t.co/Bfd3B5XhPZ> (em breve no <https://...>)


**Neto Ferraz** admin19:05:04


Confira o Tweet de @turicas:  
<https://twitter.com/turicas/status/1067161458214092802?s=09>

Twitter

Álvaro Justen

Recebemos o quadro de sócios e administradores das empresas brasileiras via LAI e eu já extraí. Dados em:  
<https://t.co/Bfd3B5XhPZ> (em breve no <https://...>)


**Julio Trecenti** sucesso19:08:05

**George Santiago** admin19:11:29

Álvaro Justen

Dados e script liberados: <https://twitter.com/turicas/status/1067...>

TOP TOP TOP TOP TOP TOP TOP TOP TOP TOP


**Julio Trecenti** uhuu20:03:06

20:03:41

```
> dplyr::glimpse(d_emp)
Observations: 100,000
Variables: 37
$ indicador_full_diario      <chr> "F", "F..."
$ tipo_atualizacao          <chr> "", "", "..."
```

**R Brasil** 1031 membros

Mostrar mensagens recentes

**George Santiago**

Olá, pessoal.

Construí um pacote para tratar a base de dados do Cadastro Nacional de Pessoas Jurídicas (CNPJ) disponibilizado pela Receita Federal.

<https://github.com/georgevbsantiago/qsacnpj>

É a primeira versão do pacote. No README fiz algumas observações sobre a execução, desempenho e resultado final.

Mais tarde, disponibilizo o link do arquivo com os dados tratados no formato do SQLite.


@BrownSantana começou a estudar se é possível implementá-lo no Spark, usando o pacote [sparklyr](#), com o objetivo de melhorar a performance. Atualmente, o pacote demora cerca de 1 hora e 25 minutos, gerando um arquivo SQLite de +/- 24Gb, usando um notebook com processador i7 5ª Geração, 16Gb DDR3 e disco HDD.

Desde já, agradeço se alguém quiser colaborar com o desenvolvimento do pacote.

GitHub

[georgevbsantiago/qsacnpj](#)

Pacote que trata e organiza os dados do Cadastro Nacional da Pessoa Jurídica (CNPJ) -



Capital técnico e capital social para controle social sobre o capital econômico

An aerial photograph taken from an airplane window, showing the wing of the aircraft in the upper right corner. Below the wing, a large body of water, likely a reservoir or lake, stretches across the middle of the frame. A bridge with a distinctive circular design crosses the water. The surrounding landscape is a mix of green fields, trees, and residential areas with red-roofed houses. The sky is filled with white clouds.

Obrigado

- Telegram: @FBarbalho
- Twitter: @barbalhofernand
- <https://medium.com/tchiluanda>

Hora de voltar para casa. Brasília 2019

Com exceção da foto de satélite as demais  
foram feitas pelo autor desta apresentação