

Oficina ASI¹/Labhinova² sobre Ciência de Dados Aplicada ao Poder Legislativo: Análise Exploratória de Dados Abertos Usando Ferramentas Livres.

Câmara Legislativa do Distrito Federal
Mesa Diretora
Vice-Presidência
Coordenadoria de Modernização e Informática

Março de 2020

¹Área de Sistema de Informação

²Laboratório Hacker de Inovação da Câmara Legislativa do Distrito Federal

- Estratégia de Sistema de Informação (ESI)
- Plano Setorial do GPI 2020: Meta 7 Ação 4

- Estratégia de Sistema de Informação (ESI)
- Plano Setorial do GPI 2020: Meta 7 Ação 4

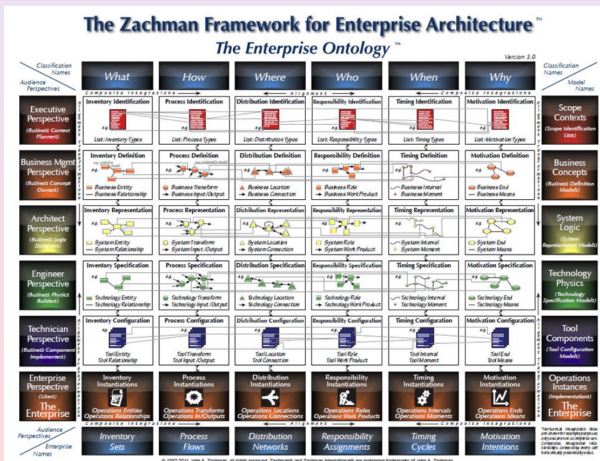


Figura 4. Estrutura Zachman para Arquitetura Organizacional.

Fonte: ZACHMAN, 2018.

Sumário

- TODO
- 1 Conceitos Básicos
 - Ciência de Dados
 - Análise Exploratória de Dados
 - Dados Abertos
 - Ferramentas Livres
- 2 Domínio de Aplicação
 - Direitos Humanos
 - Feminicídio
 - Auxílio Natalidade
 - Bolsa Família
- 3 Computação
 - Fontes de Dados Abertos
 - Formatos de Dados Abertos
 - Formas de Acesso aos Dados
 - Manipulação dos Dados
- 4 Estatística

AFAZERES

Afazeres

Estou anotando afazeres.

Esses slides serão removidos na versão final.

- ⇒ Remover anotações ao final alterando a opção no arquivo `cfg.tex` que vai esconder as notas automaticamente.
- ⇒ Remover esse `tex` comentando a linha em `dev.tex`
- ⇒ Colocar fonte do diagrama venn nas referencias;
- ⇒ Adicionar outras fontes de dados abertos. Os dados da Amazon, Yahoo, e bibliotecas python que entregam dados para as pessoas praticarem;
- Pegar Vídeo de Evolução do Uso da Linguagem Python do Stackoverflow para exibir no dia;
- Colocar o trecho do livro do Data Science do zero em algum lugar.
- Copiar a versão final do pdf para a pasta do github

Conceitos Básicos: Ciência de Dados

Ciência de Dados Aplicado ao Poder Legislativo

"Ciência de Dados Aplicada ao Poder Legislativo: Análise Exploratória de Dados Abertos Usando Ferramentas Livres"

Ciência de Dados

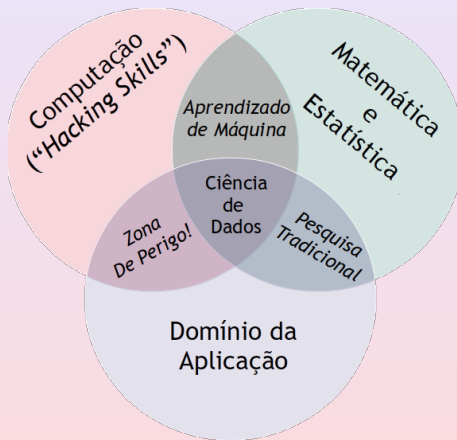


Figura: Diagrama Venn da Ciência de Dados

└ Conceitos Básicos

└ Ciência de Dados

└ Ciência de Dados



Figura: Diagrama Venn da Ciência de Dados

- ⇒ Ciência de Dados está na **intersecção** da ciência da computação, estatística e domínios de aplicação reais.
- ⇒ Da ciência da computação vem tecnologias de aprendizado de máquina e de computação de alto desempenho para lidar com escala. Em termos simples, estamos falando de hacking skills, ou seja habilidades de computação necessárias para **automatizar** a manipulação dos dados.
- ⇒ Da estatística, vem uma longa tradição de análise exploratória de dados, teste de significância e visualização.
- ⇒ Dos domínios de aplicações e ciências vêm desafios dignos de batalha e padrões de avaliação para avaliar quando eles foram adequadamente conquistados.

Conceitos Básicos: Análise Exploratória de Dados

Análise Exploratória de Dados

"Ciência de Dados Aplicada ao Poder Legislativo: **Análise Exploratória de Dados** Abertos Usando Ferramentas Livres"

Análise Exploratória de Dados

Conceito

A Análise Exploratória de Dados ou EDA (*Exploratory Data Analysis*) consiste em analisar bases de dados e extrair informações úteis dos dados através de técnicas de visualizações. Assim o cientista de dados é capaz de formular algumas hipóteses a cerca das informações que estão a sua disposição.

Conceitos Básicos: Dados Abertos

Dados Abertos

"Ciência de Dados Aplicada ao Poder Legislativo: Análise Exploratória de **Dados Abertos** Usando Ferramentas Livres"

Dados Abertos

Conceito

De acordo com a organização Open Definition:

"dados são abertos quando qualquer pessoa pode livremente acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito a, no máximo, a exigências que visem preservar sua proveniência e sua abertura."

Mais informações

<http://www.dados.gov.br/pagina/dados-abertos>

Conceitos Básicos: Ferramentas Livres

Ferramentas Livres

"Ciência de Dados Aplicada ao Poder Legislativo: Análise Exploratória de Dados Abertos Usando **Ferramentas Livres**"

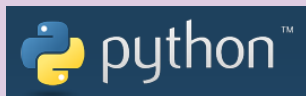
Ferramentas Livres

Software Livre

Software livre é o software que concede liberdade ao usuário para executar, acessar e modificar o código fonte, e redistribuir cópias com ou sem modificações. Sua definição é estabelecida pela Free Software Foundation (FSF) em conjunto com o projeto GNU.

Ferramentas Livres para Ciência de Dados

Linguagem de Programação



"Python is powerful... and fast; plays well with others; runs everywhere; is friendly & easy to learn; is Open."

Mais informações

<https://www.python.org/about/>

Ferramentas Livres para Ciência de Dados

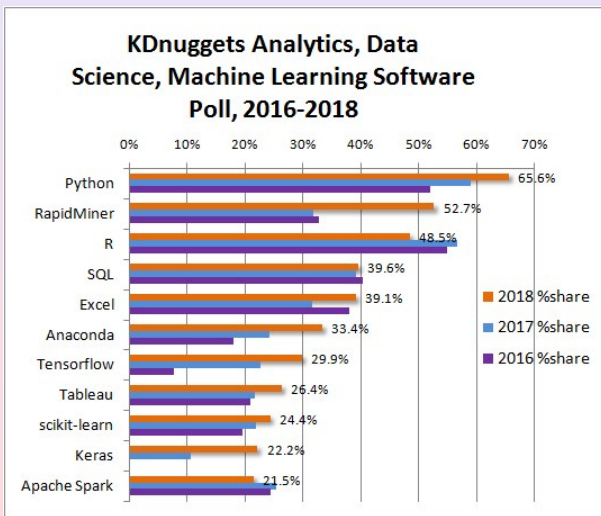


Figura: Por que Python?

└─ Conceitos Básicos

└─ Ferramentas Livres

└─ Ferramentas Livres para Ciência de Dados

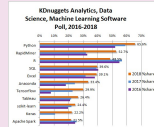


Figura: Por que Python?

- Data science involves extrapolating useful information from massive stores of statistics, registers, and data. These data are usually unsorted and difficult to correlate with any meaningful accuracy. Machine learning can make connections between disparate datasets but requires serious computational sophistry and power.
- Python fills this need by being a general-purpose programming language. It allows you to create CSV output for easy data reading in a spreadsheet. Alternatively, more complicated file outputs that can be ingested by machine learning clusters for computation.
- There are now over 70,000 libraries in the Python Package Index, and that number continues to grow. As previously mentioned, Python offers many libraries geared toward data science. A simple Google search reveals plenty of Top 10 Python libraries for data science lists. Arguably, the most popular data analysis library is an open source library called pandas. It is a high-performance set of applications that make data analysis in Python a much simpler task

Ferramentas Livres para Ciência de Dados

Análise de Dados

- Dask
- pandas
- NumPY
- Numba

Visualização

- Matplotlib
- Datashader
- Bokeh
- Holoviews

Machine & Deep Learning

- scikit-learn
- Theano
- Tensorflow

Nessa oficina...

Para estudar, praticar e aprender



"The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text (...)"

Jupyter

<https://jupyter.org>

Nessa oficina...

Para fazer Análise de Dados



"pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language."

Pandas

<https://pandas.pydata.org>

Nessa oficina...

Para visualizar



"Bokeh is an interactive visualization library for modern web browsers. It provides elegant, concise construction of versatile graphics, and affords high-performance interactivity over large or streaming datasets (...)"

Bokeh

<https://docs.bokeh.org>

Domínio de Aplicação

DOMÍNIO DE APLICAÇÃO

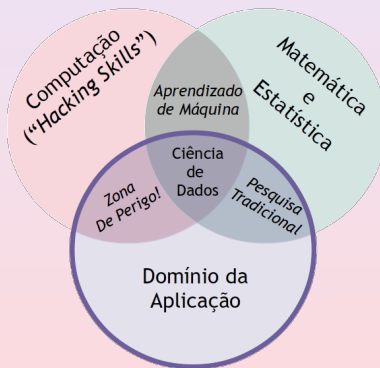


Figura: Ciência de Dados: Ênfase no Domínio da Aplicação

Direitos Humanos

Conceito

Falar sobre o Domínio da Aplicação

Direitos Humanos

Conceito

Falar sobre o Domínio da Aplicação

Direitos Humanos

Conceito

Falar sobre o Domínio da Aplicação

Direitos Humanos

Conceito

Falar sobre o Domínio da Aplicação

Computação

Computação ("Hacking Skills")

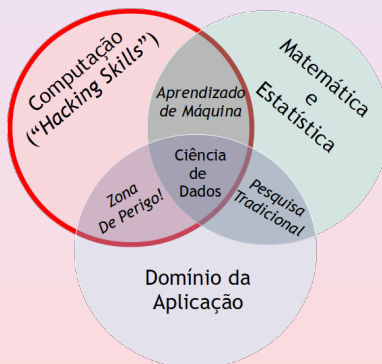


Figura: Ciência de Dados: Ênfase em Computação

Onde estão os dados?



(a) Portal Brasileiro de Dados Abertos



(b) Portal da Transparência
Controladoria-Geral Da União

(c) Portal da Transparência
do Distrito Federal

(d) GOV DATA



(e) Senado Federal



(f) Câmara
dos Deputados



(g) Câmara Legislativa
do Distrito Federal

Figura: Portais de Dados Abertos

Onde estão os dados?

Outras fontes

Colocar outras fontes

Onde estão os dados?

Portal Brasileiro de Dados Abertos



Figura: Portal Brasileiro de Dados Abertos

É um catálogo!

O portal funciona como um catálogo federado que facilita a busca e uso de dados publicados pelos órgãos do governo.

2020-03-03

ASI

└─ Computação

└─ Fontes de Dados Abertos

└─ Onde estão os dados?

Onde estão os dados?



É um catálogo!

O portal funciona como um catálogo federado que facilita a busca e uso de dados publicados pelos órgãos do governo.

Falar que é importante apresentar, pelo menos, o Portal Brasileiro de Dados Abertos.

Explorar o Portal Brasileiro de Dados Abertos

Passo-a-passo

- Abrir o “Portal Brasileiro de Dados Abertos”;
- Pesquisar por “Feminicídio” como um exemplo de busca;
- Acessar “Violência contra a mulher no DF”;
- Entender a página que é aberta;
- Acessar “Crimes de Feminicídio no DF”;
- Clicar em “Ir para recurso”;
- Abrir a planilha e mostrar um exemplo de algo que se encontra no portal. Trata-se de um relatório em formato de planilha excel;
- Fechar a planilha e retornar para o portal;

Explorar o Portal Brasileiro de Dados Abertos

Passo-a-passo (Cont...)

- Na coluna dos filtros, atentar-se para os “Formatos” de dados abertos mais comuns. Nessa oficina vamos apresentar:
 - **CSV;**
 - **JSON;**
- Além disso, também é necessário entender as formas de acesso aos dados que podem se dar:
 - ❶ **Manual fazendo download de um arquivo** - O Cientista de Dados faz o download dos dados e depois utiliza um software para auxiliá-lo a analisar os dados;
 - ❷ **Acesso por meio de uma API** - O Cientista de Dados acessa os dados via programação usando uma *Application Programming Interface* fornecida pela fonte dos dados.

⇒ Na sequência, vamos apresentar os formatos CSV e JSON, que geralmente são disponibilizados para download e, em seguida, experimentar o acesso utilizando uma API.

Principais Formatos: CSV

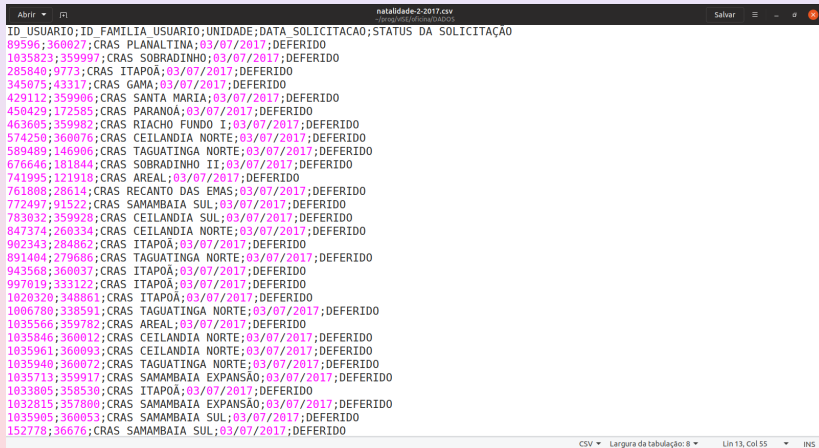
CSV

Por definição, CSV é um formato de arquivo que significa “comma-separated-values” (valores separados por vírgulas). Isso significa que os campos de dados indicados neste formato normalmente são separados ou delimitados por uma vírgula.

Exemplo

- 1 Voltar para o “[Portal Brasileiro de Dados Abertos](#)”;
- 2 Ir na aba “[Organizações](#)” e na segunda página abrir “[Distrito Federal](#)”;
- 3 Na barra de pesquisas pesquisar por “[Direitos Humanos](#)”;
- 4 Abrir “[Benefício eventual em virtude de nascimento \(Auxílio Natalidade\)](#)”;
- 5 Ir para o recurso “[2º semestre/2017](#)”;

Principais Formatos: CSV



```
Abzir  [F] natalidade-2-2017.csv
~prog/vista/office/DADOS

ID USUARIO;ID FAMILIA USUARIO;UNIDADE;DATA SOLICITACAO;STATUS DA SOLICITACAO
89596;360027;CRAS PLANALTIMA;03/07/2017;DEFERIDO
1035823;359997;CRAS SOBRADINHO;03/07/2017;DEFERIDO
285840;9773;CRAS ITAPOA;03/07/2017;DEFERIDO
345075;43317;CRAS GAMA;03/07/2017;DEFERIDO
429112;359906;CRAS SANTA MARIA;03/07/2017;DEFERIDO
450429;172585;CRAS PARANOA;03/07/2017;DEFERIDO
463605;359982;CRAS RIACHO FUNDO I;03/07/2017;DEFERIDO
574250;360076;CRAS CEILANDIA NORTE;03/07/2017;DEFERIDO
589489;146906;CRAS TAGUATINGA NORTE;03/07/2017;DEFERIDO
676646;181844;CRAS SOBRADINHO II;03/07/2017;DEFERIDO
741995;121918;CRAS AREAL;03/07/2017;DEFERIDO
761808;28614;CRAS RECANTO DAS EMAS;03/07/2017;DEFERIDO
772497;91522;CRAS SAMAMBAIA SUL;03/07/2017;DEFERIDO
783032;359928;CRAS CEILANDIA SUL;03/07/2017;DEFERIDO
847374;260334;CRAS CEILANDIA NORTE;03/07/2017;DEFERIDO
902343;284862;CRAS ITAPOA;03/07/2017;DEFERIDO
891404;279686;CRAS TAGUATINGA NORTE;03/07/2017;DEFERIDO
943568;360037;CRAS ITAPOA;03/07/2017;DEFERIDO
997019;333122;CRAS ITAPOA;03/07/2017;DEFERIDO
1020320;348861;CRAS ITAPOA;03/07/2017;DEFERIDO
1006780;338591;CRAS TAGUATINGA NORTE;03/07/2017;DEFERIDO
1035566;359782;CRAS AREAL;03/07/2017;DEFERIDO
1035846;360012;CRAS CEILANDIA NORTE;03/07/2017;DEFERIDO
1035961;360093;CRAS CEILANDIA NORTE;03/07/2017;DEFERIDO
1035940;360072;CRAS TAGUATINGA NORTE;03/07/2017;DEFERIDO
1035713;359917;CRAS SAMAMBAIA EXPANSAO;03/07/2017;DEFERIDO
1033805;358530;CRAS ITAPOA;03/07/2017;DEFERIDO
1032815;357800;CRAS SAMAMBAIA EXPANSAO;03/07/2017;DEFERIDO
1035905;360053;CRAS SAMAMBAIA SUL;03/07/2017;DEFERIDO
152778;36676;CRAS SAMAMBAIA SUL;03/07/2017;DEFERIDO
```

CSV Largura da tabulação: 8 Lin 13, Col 55 INS

Figura: Exemplo de arquivo CSV

Principais Formatos: JSON

JSON

JSON (JavaScript Object Notation - Notação de Objetos JavaScript) é uma formatação leve de troca de dados. Para seres humanos, é fácil de ler e escrever. Para máquinas, é fácil de interpretar e gerar.

Exemplo

- 1 Acessar “Dados Bolsa Família para Brasília para Janeiro de 2020”;

Principais Formatos: JSON

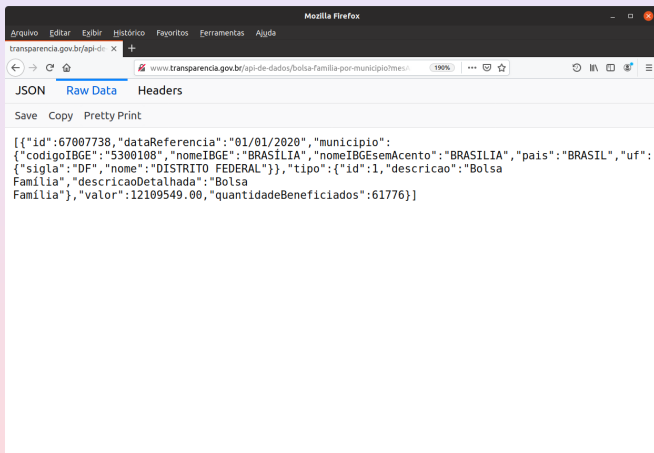


Figura: JSON RAW

Principais Formatos: JSON

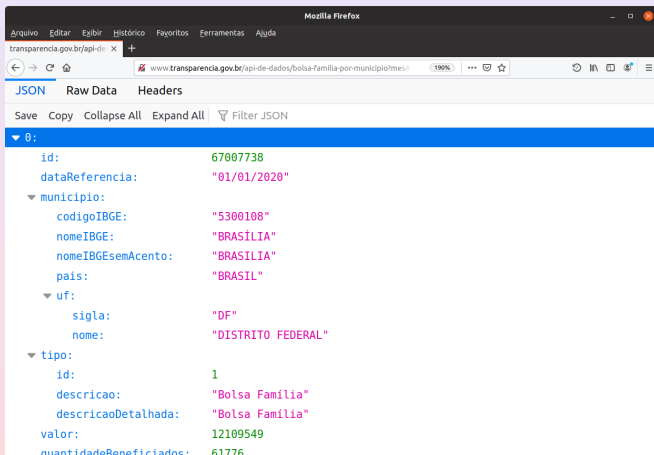


Figura: JSON formatado e colorido pelo browser Mozilla Firefox

Manipulação dos Dados



Figura: Mão na massa!

Manipulação dos Dados

Acessar:

<https://mybinder.org/v2/gh/Isegoria/oficina/master?filepath=apresentacao.ipynb>



Figura: Copie o link acima ou use o qr-code!

Vamos continuar a oficina por lá!

Manipulação dos Dados: QR-CODE MAIOR



Figura: QR-CODE MAIOR

Estatística

Matemática e Estatística

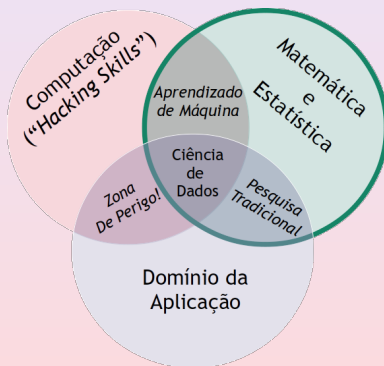


Figura: Ciência de Dados: Ênfase em Matemática e Estatística

Considerações Finais

"Soterrados sob os dados, estão as respostas para as inúmeras questões que ninguém nunca pensou em perguntar. Um dos objetivos da Ciência de Dados é encontrá-las."

(Joel Grus, Data Science do Zero, adaptado)

Questões

- Por que o total de beneficiários varia mês a mês?

Bibliografia I



CHEN, Daniel Y.

Análise de dados com Python e Pandas.

São Paulo: Novatec, 2018.



MOORE, David S.

A Estatística básica e sua prática.

3. ed. Rio de Janeiro: LTC, 2005.



ROSEN, Kenneth H.

Discrete mathematics and its applications.

7. ed. New York: McGraw-Hill, 2012.



SATO, K.

Levy Processes and Infinitely Divisible Distributions.

Cambridge: Cambridge University Press, 1999.

Bibliografia II



SKIENA, Steve S.

The data science design manual..

Stony Brook: Springer-Verlag, 2017.



CONWAY, D.

The data science venn diagram

Disponível em:

<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

Acesso em: 3 mar 2020.



ECMA INTERNATIONAL.

The JSON data interchange syntax..

Disponível em:

<https://www.ecma-international.org/publicationa/standards/Ecma-404>

Acesso em: 11 fev 2020.

Bibliografia III



INSTITUTO DE MATEMÁTICA E ESTATÍSTICA,
UNIVERSIDADE DE SÃO PAULO.

Introdução à ciência da computação com python..

Disponível em:

[https://www.youtube.com/watch?v=WT_zCgSHSTQlist =
PLcoJJSvnDgcKpOi_UeneTNTIVOigRQwcn](https://www.youtube.com/watch?v=WT_zCgSHSTQlist=PLcoJJSvnDgcKpOi_UeneTNTIVOigRQwcn). Acesso em :
12 fev 2020.



Protter, P.

Stochastic integration and differential equations.

U.S. Government Printing Office, 2004.

Bibliografia IV



CONWAY, D.

CSE 519 - data science fall 19

Disponível em:

<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>.

Acesso em: 3 mar 2020.