
Y
X
ESCOLA
DE
DADOS

Dados, uma fonte a ser entrevistada

Adriano Belisário
belisario@ok.org.br

O que iremos aprender?

- O que são dados abertos e dados legíveis por máquinas;
- O que é e como abrir um arquivo CSV;
- Quais são os tipos de dados mais comuns e como configurá-los;
- Operações básicas: ordenar e filtrar dados por diferentes critérios;
- Agrupando informações: como usar tabela dinâmica para analisar dados;
- Análise de dados com taxas e medidas de tendência central;
- Como cruzar dados;

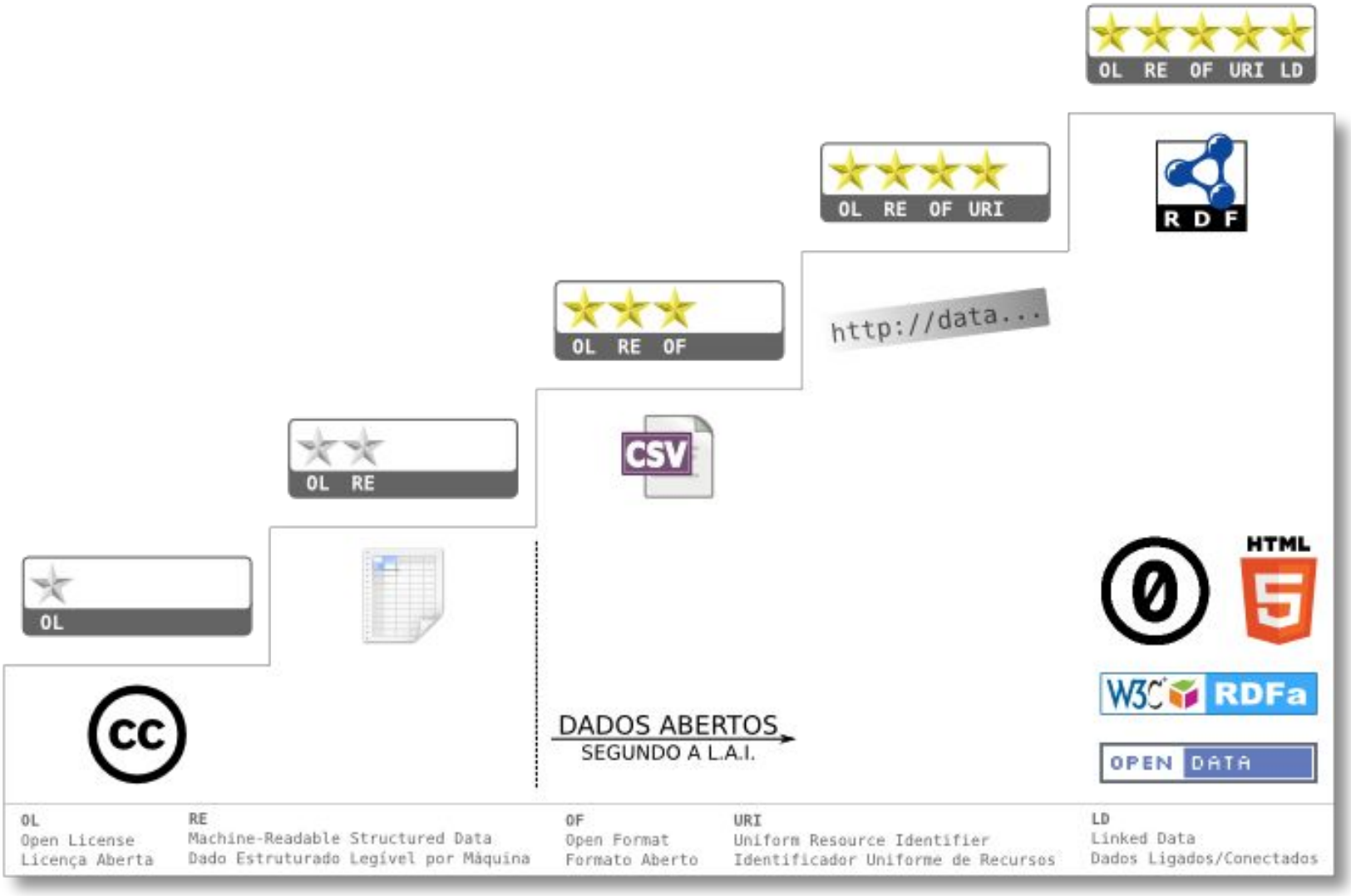
O que são dados abertos e legíveis por máquinas?

Licenças Abertas

O que isso significa na prática?

"Aberto significa que qualquer um pode livre e gratuitamente acessar, usar, modificar e compartilhar para qualquer propósito (sujeito, no máximo, à exigência de preservação da proveniência e abertura)"

opendefinition.org



https://br.okfn.org/2013/01/17/maturidade-em-dados-abertos-entenda-as-5-estrelas/

O que é e como abrir um CSV?

O que é um CSV?

Comma Separated Values

É um arquivo de texto, onde as colunas são separadas por um **delimitador**, já as linhas por quebras de linha.

Vírgula e ponto-e-vírgula são os delimitadores mais comuns.

Sem limite de linhas.

Dicas para abrir um CSV

E evitar problemas futuros

- Prefira a opção “Importar” à de “Abrir”;
- Verifique o delimitador utilizado pelo arquivo. Se as colunas não abrirem corretamente, tente importar novamente usando outro delimitador;
- Confira se a “localização” do Google Sheets condiz com o formato dos dados;
- Evite usar a conversão automática de “tipos de dados”. Configure-os manualmente.

Quais são os tipos de dados mais comuns e como configurá-los?

Tipos de Dados

“Data Types” comuns

- **Datas**: DD/MM/YYYY; MM/DD/YYYY
- **Caracteres/strings**
- **Números inteiros**
- **Números “quebrados” (float)**
- **NULL (nulo/vazio)**

Mas o que pode dar errado?

Imagine um CNPJ

00328442900119

Mas o que pode dar errado?

Números não precisam de zero à esquerda

00328442900119

Operações básicas
ordenar e filtrar dados
por diferentes critérios

Operações Básicas

Filtrando e ordenando

- Subset: filtros básicos/categóricos e filtros condicionais;
- Sort by: ordenando intervalos e ordenando páginas;

Tabela Dinâmica

Tabela Dinâmica

Entrevistando seus dados

- É uma forma de criar uma nova tabela, que utilizará seus dados para responder a perguntas;

Tabela Dinâmica

Entrevistando seus dados

- É uma forma de criar uma nova tabela, que utilizará seus dados para responder a perguntas;
- Define a pergunta e defina quais colunas estão implicadas nela;

Tabela Dinâmica

Entrevistando seus dados

- É uma forma de criar uma nova tabela, que utilizará seus dados para responder a perguntas;
- Define a pergunta e defina quais colunas estão implicadas nela;
- Selecione o intervalo desejado ou toda planilha e crie a tabela dinâmica;

Tabela Dinâmica

Entrevistando seus dados

- É uma forma de criar uma nova tabela, que utilizará seus dados para responder a perguntas;
- Define a pergunta e defina quais colunas estão implicadas nela;
- Selecione o intervalo desejado ou toda planilha e crie a tabela dinâmica;
- Configure os campos de “linhas”, “colunas” e “valores”, de acordo com sua pergunta;

Funções da tabela dinâmica

- SUM → Soma
- COUNTA → Conta todos os registros, inclusive itens duplicados ou em branco
- COUNT → Conta todos os registros numéricos
- COUNTUNIQUE → Conta apenas os registros únicos (bom para contar categorias)
- AVERAGE → Média
- MAX → Valor máximo
- MIN → Valor mínimo
- MEDIAN → Mediana
- PRODUCT → Multiplicação
- STDEV → Desvio padrão para amostras
- STDEVP → Desvio padrão para populações
- VAR → Variância para amostras
- VARP → Variância para populações

Medidas de tendência central

Olho nos outliers

Valores atípicos

Nome	Idade
Paulo	3
Julia	10
Ricardo	209
Samuel	234
Samara	25

Outliers

Podem distorcer sua média

Imagine que você foi chamado para trabalhar em uma empresa com média salarial de R$ 31 mil por mês. Parece bom, não?

Mas lembre-se: a média pode ser enganosa se tivermos outliers.

Outliers

Podem distorcer sua média

Empresa Xtreme

ID	Cargo	Salário mensal
1	Presidente	200.000
2	Gerente de vendas	5.000
3	Gerente de produção	5.000
4	Administrador	4.000
5	Vendedor	3.000
6	Secretário	2.000
7	Faxineiro	1.000

Mediana

A irmã menos famosa da média

Empresa Xtreme

ID	Cargo	Salário mensal
1	Presidente	200.000
2	Gerente de vendas	5.000
3	Gerente de produção	5.000
4	Administrador	4.000
5	Vendedor	3.000
6	Secretário	2.000
7	Faxineiro	1.000

Mediana

A irmã menos famosa da média

Empresa Ygual

ID	Cargo	Salário mensal
1	Presidente	200.000
2	Gerente de vendas	5.000
3	Gerente de produção	5.000
4	Gerente de mídia	5.000
5	Administrador	4.000
6	Vendedor	3.000
7	Secretário	2.000
8	Faxineiro	1.000

Moda

O valor mais recorrente

Empresa Xtreme

ID	Cargo	Salário mensal
1	Presidente	200.000
2	Gerente de vendas	5.000
3	Gerente de produção	5.000
4	Administrador	4.000
5	Vendedor	3.000
6	Secretário	2.000
7	Faxineiro	1.000

Taxas

Variação percentual

Para comparar números

- Diminuir o VALOR pelo VALOR DE REFERÊNCIA;
- Dividir o resultado da etapa anterior pelo VALOR DE REFERÊNCIA;
- Transformar em percentagem multiplicando o resultado da etapa anterior por 100;
- Exemplo: o PIB mundial foi de USD 85.798 trilhões em 2018 e USD 80.886 tri em 2017: qual a variação percentual no período?

Variação percentual

Para comparar números

(Ano de 2018) - (Ano de 2017)
85.798 - 80.886 = 4.912
-
(Dividimos o resultado pelo valor de referência)
4.912 / 80.886 = 0.06072744356
-
(Em formato percentual)
0.06072744356*100 = 6.0727443562545
6.07%

A parte e o todo

A percentagem é sua amiga

- Use a regra de três para descobrir a proporção em % de um determinado valor em relação a outro.

- Exemplo: o PIB mundial foi de R$ 85.791 trilhões em 2018, o Brasil registrou R$ 1.869 trilhões.

- Qual a participação do PIB brasileiro no PIB mundial?

A parte e o todo

Para comparar números

X ---------- 1.869
100 ----- 85.791

X/100 = 1.869/85.791

X/100 = 0.02178550197

X = 0.02178550197*100

2.17%

Taxas

Nem sempre a percentagem resolve

- Para comparar fenômenos em populações grandes;
- Homicídios em geral são expressos considerando a taxa por cem mil habitantes;
- Fórmula: EVENTOS / POPULAÇÃO * UNIDADE

- Exemplo: considerando uma população de 208.494.900 de pessoas e 51.589 homicídios no Brasil por um lado e 131.788.270 de pessoas e 33.341 homicídios no México. Calcule a taxa de homicídios por cem mil habitantes nos dois países.

Taxas

51.589/208.494.900 = 0.0002474353

0.0002474353 * 100000 = 24.74353

33.341 / 131.788.270 = 0.00025298913

0.00025298913 * 100000 = 25.298913

Taxas

24 hom. por 100 mil/hab.

25 hom. por 100 mil/hab

Como cruzar dados?

Cruzando dados

Duas tabelas e um identificador

Vamos importar as duas tabelas para o mesmo arquivo e utilizar um campo em comum para puxar informações de uma para outra.

PROCV/VLOOKUP

Procura vertical

A função é composta por quatro parâmetros no Google Sheets, que são listados entre parênteses.

=PROCV(D2;A2:B5;2;FALSO)

ATENÇÃO: A depender da localização configurada no seu Google Sheets (Arquivo > Configurações da planilha), o nome da função (PROCV ou VLOOKUP) e o separador usado entre os parâmetros (ponto-e-vírgula ou vírgula) podem variar.

PROCV/VLOOKUP

Procura vertical

=PROCV(D2;A2:B5;2;FALSO)

1) O primeiro corresponde ao valor a ser buscado.

2) O segundo corresponde ao intervalo onde será feita a busca, sendo que a primeira coluna deve utilizar o mesmo identificador da coluna especificada no primeiro parâmetro;

3) No intervalo especificado no item anterior, identificamos qual a posição da coluna que queremos retornar.

4) Em geral, usamos sempre “FALSO”.

INDEX MATCH

Outra opção, mais flexível

Primeiro passo: =MATCH(search_key, range, [search_type])