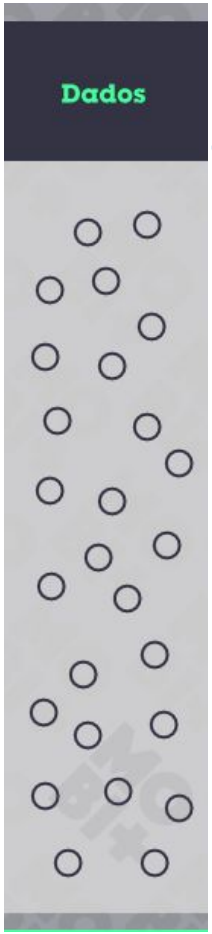
Ciência dos dados: a profissão do momento

Bárbara Silveira
Karen Martins
Larissa Maia

- O que é um dado?
- E informação?
- E conhecimento?

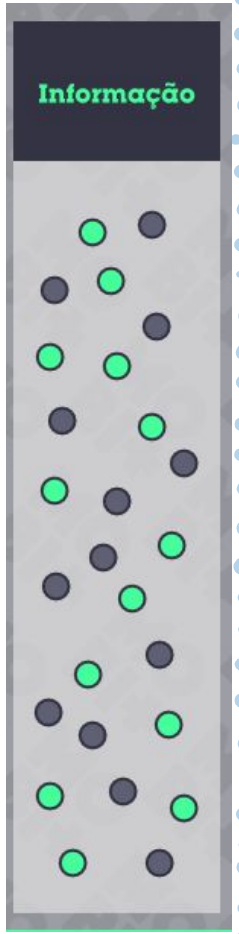
Dado

- Conjunto de valores ou ocorrências em um estado bruto, que ainda não passou por nenhum processo e nenhuma organização para ser utilizado
- Por exemplo: Um pingo de chuva. Quando você está caminhando na rua e sente um pingo caindo em você
- Aquele fato não representa que está chovendo, pode significar um ar condicionado, um pássaro ou qualquer outra fonte estranha



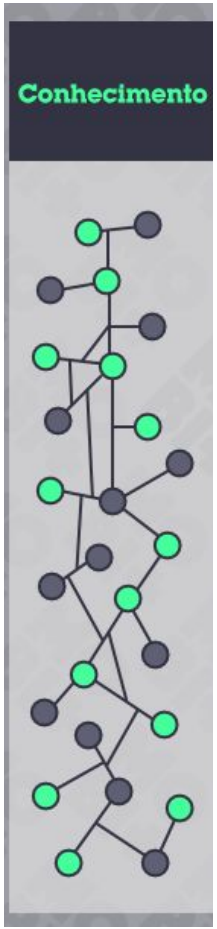
Informação

- É o dado já processado, no qual já teve algum tipo de organização e será utilizado para qualquer tipo de conceito tanto para qualitativo ou quantitativo
- Por exemplo: É a descoberta que vai chover
  - Você olha para o céu e começa a perceber que existem nuvens escuras (outro dado), além disso percebe que está pingando mais vezes
  - Então, nesse momento, a partir de vários dados você chega a uma conclusão: Vai chover



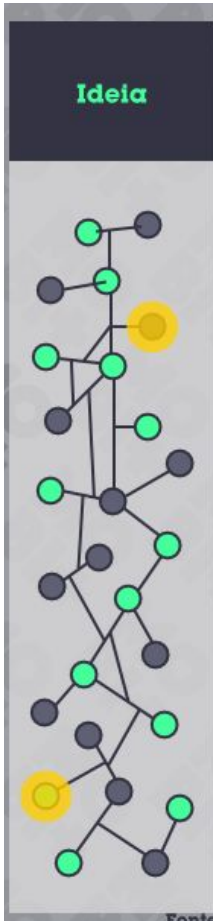
Conhecimento

- Quando você, a partir das informações que tem, pode prever o que vai acontecer, seja pelo **histórico dos fatos anteriores** ou por **novas conclusões**, está utilizando o seu conhecimento
- Por exemplo:
  - Vou me molhar
  - Com a chuva vou ficar molhado, não poderei trabalhar molhado e não conseguirei entregar o relatório



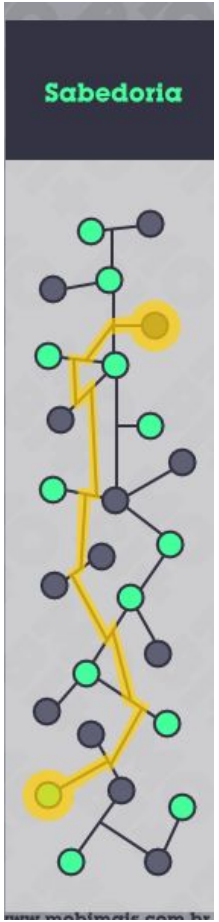
Ideia

- Soluções para esse problema são as ideias que você vai ter
- Por exemplo:
  - Comprar um guarda-chuva ou pedir um táxi no aplicativo?



Sabedoria

- A sabedoria é o que vai fazer com tudo isso. Qual a melhor decisão a ser tomada
- Por exemplo:
  - Se você vai pegar o taxi ou comprar o guarda-chuva
  - Usando seu conhecimento:
    - Os guarda-chuvas custam R$ 15.00 (conhecimento), e o táxi até meu escritório custa R$ 10.00.

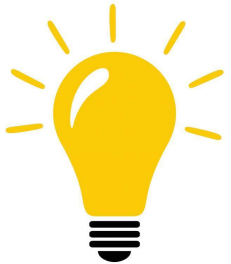


Quantidade de dados gerados pela internet

- Você já parou para pensar como a Netflix consegue "adivinhar" os filmes que você gostaria de assistir?
- Como o Youtube possui vários vídeos interessantes como recomendação?
- E aquele site de compras que parece saber todas as coisas que você gostaria de comprar?

Ciência dos dados

- Existe uma área inteira na Computação a qual estuda como **gerenciar**, **analisar** e **visualizar** essa quantidade enorme de dados
- Com isso, é possível **produzir conhecimentos** para auxiliar a responder todas essas perguntas
- Obtendo insights que podem ajudar as organizações a **melhorar** seus negócios

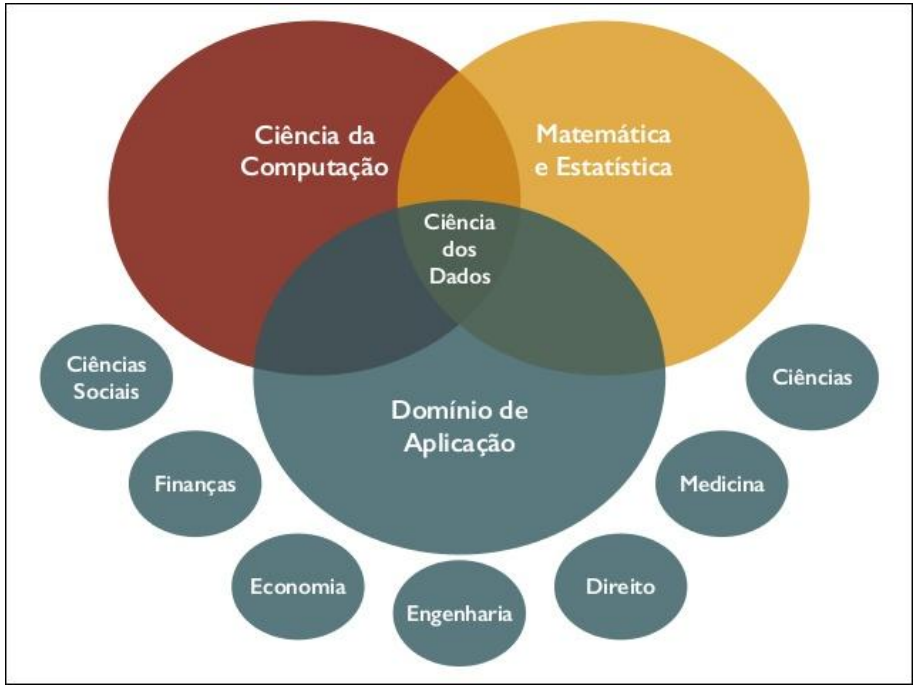


Ciência dos dados

- Definição:
  - “Ciência dos dados é a ciência que estuda formas de **extrair conhecimento** generalizável **a partir de dados**” - Vasant Dhar

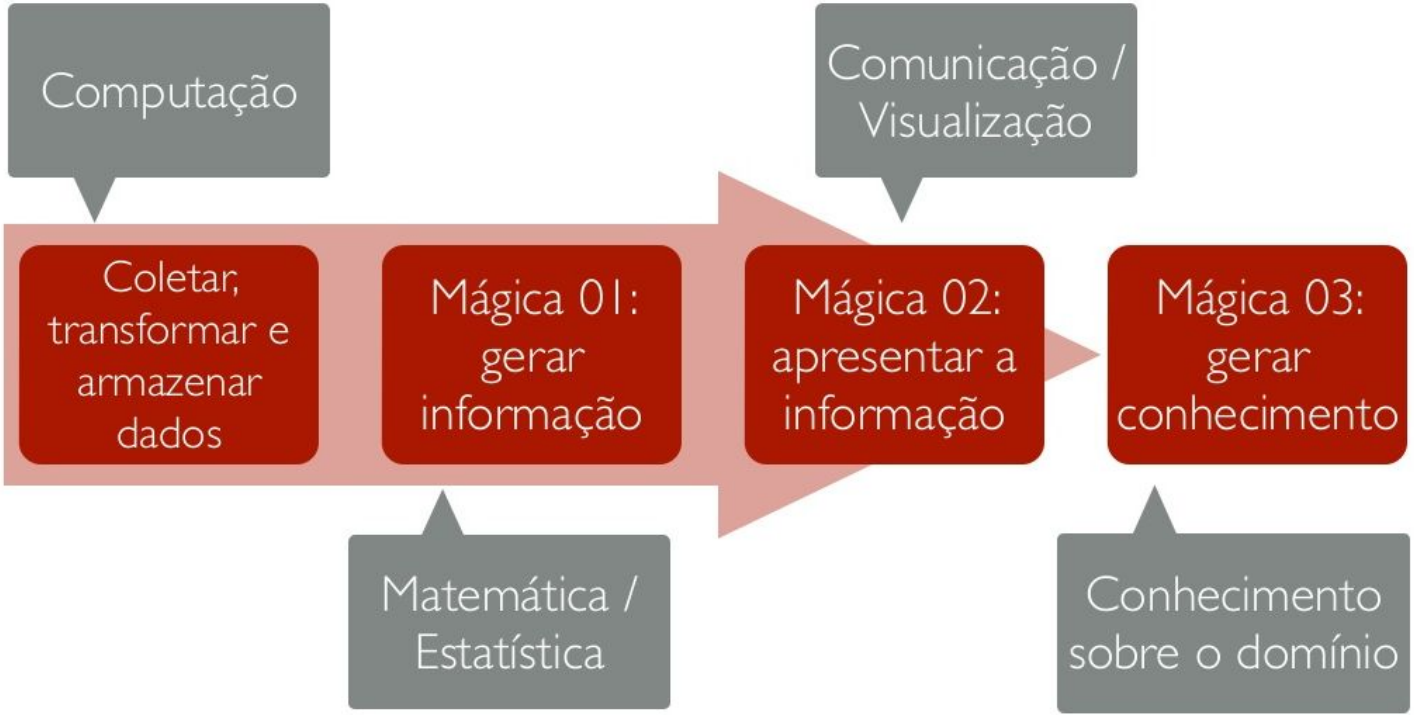
Ciência dos dados

- É uma área multidisciplinar



Ciência dos dados

- Definição:



Ciência dos dados

- Cientista de dados: utilizar **dados do presente e do passado** para criar modelos que possam **prever comportamentos futuros**
- Aqueles que utilizam a tecnologia a seu favor tendem a tomar **decisões mais rápidas e eficazes**, aumentando a lucratividade
- A abordagem para a análise depende da indústria e das necessidades específicas do negócio
- Muitos projetos reúnem pessoas de outras áreas (gerente da empresa)

Exemplos

?

A natureza humana é pessimista ou otimista?

Pesquisa

- Desde 1969 teorias foram discutidas mas nenhuma foi conclusiva
- 10 línguas analisadas
- Livros, sites e letras de músicas
- Somente do twitter foram 100 bilhões de palavras
- Panometer.org

Fluxo

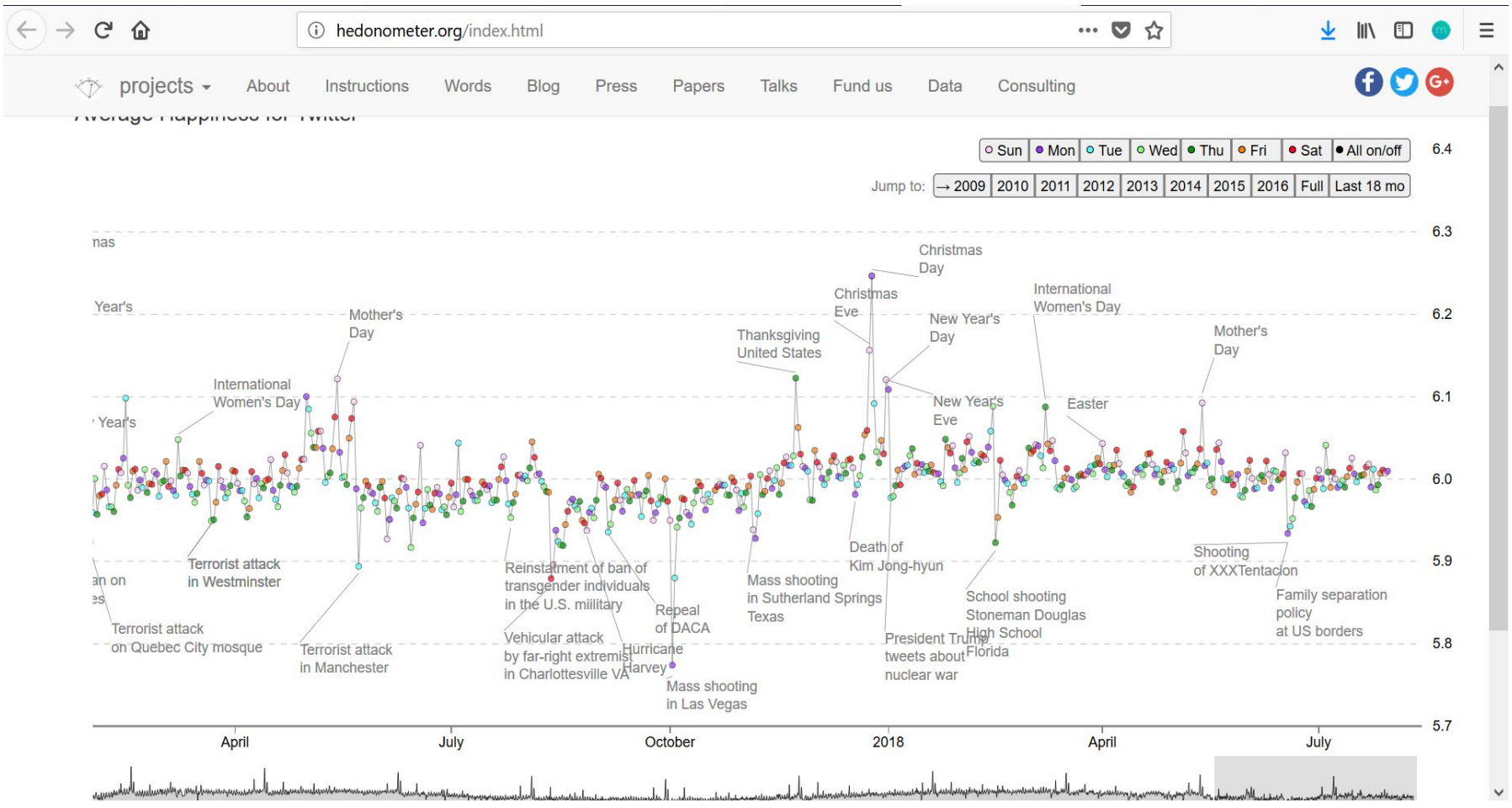
1º Coleta dos dados

2º Identificação das 10 mil palavras mais utilizadas

3º Palavras foram ranqueadas de acordo com uma escala

4º Cada palavra teve sua média calculada

5º Descoberta: pessoas tendem a usar *happy words*


hedonometer.org/index.html
projects
About
Instructions
Words
Blog
Press
Papers
Talks
Fund us
Data
Consulting
Sun
Mon
Tue
Wed
Thu
Fri
Sat
All on/off
Jump to:
2009
2010
2011
2012
2013
2014
2015
2016
Full
Last 18 mo
6.4
6.3
6.2
6.1
6.0
5.9
5.8
5.7
Year's
Year's
International Women's Day
Mother's Day
Terrorist attack in Westminster
Terrorist attack on Quebec City mosque
Terrorist attack in Manchester
Reinstatment of ban of transgender individuals in the U.S. military
Vehicular attack by far-right extremist in Charlottesville VA
Hurricane Harvey
Repeal of DACA
Mass shooting in Las Vegas
Thanksgiving United States
Christmas Eve
Christmas Day
New Year's Day
Mass shooting in Sutherland Springs Texas
Death of Kim Jong-hyun
President Trump tweets about nuclear war
New Year's Eve
International Women's Day
Easter
School shooting Stoneman Douglas High School Florida
Mother's Day
Shooting of XXXTentacion
Family separation policy at US borders
April
July
October
2018
April
July

Word	Happiness Rank	Average	Standard Deviation	Twitter	Google Books	New York Times	Lyrics
laughter	1	8.50	0.93	3600			1728
happiness	2	8.44	0.97	1853	2458		1230
love	3	8.42	1.10	25	317	328	23
happy	4	8.30	0.99	65	1372	1313	375
laughed	5	8.26	1.15	3334	3542		2332
laugh	6	8.22	1.37	1002	3998	4488	647
laughing	7	8.20	1.10	1579			1122
excellent	8	8.18	1.10	1496	1756	3155	
laughs	9	8.18	1.15	3554			2856
joy	10	8.16	1.05	988	2336	2723	809
successful	11	8.16	1.07	2176	1198	1565	
win	12	8.12	1.08	154	3031	776	694
rainbow	13	8.10	0.99	2726			1723
smile	14	8.10	1.01	925	2666	2898	349
won	15	8.10	1.21	810	1167	439	1493
pleasure	16	8.08	0.96	1497	1526	4253	1398
smiled	17	8.08	1.06		3537		2248
rainbows	18	8.06	1.36				4216
winning	19	8.04	1.04	1876		1426	3646
celebration	20	8.02	1.53	3306		2762	4070
enjoyed	21	8.02	1.53	1530	2908	3502	

Futuro

- Disponibilizar uma ferramenta para uso de empresas, negócios e população em geral
- Ferramentas do site somente em inglês
- Ainda está em desenvolvimento
- Prevenção de suicídios

?

O que a ciência dos dados pode te ensinar sobre o amor?

Pesquisa

- Cientista que estuda dados de relacionamento virtual
- Analisou 5.500 emails trocados com o namorado
- Em média 4 emails por dia

Descobertas

- Ela enviava mais e-mails do que ele
- Ele usou mais as palavras “me ligue” e “telefone”
- Ela é mais agressiva e obscena
- Ele tende a falar mais “não tenho certeza” e é responsável por 60% dos “me desculpa”
- Ele é quadri linguístico
- Ela usa palavras de estatística

Dados Abertos

Dados abertos

- Dados abertos são dados que qualquer pessoa pode acessar, usar e compartilhar
- Governo Aberto: https://www.youtube.com/watch?v=T6_AsumMFm4

Exemplos de Dados abertos

Para Onde Foi Meu Dinheiro?

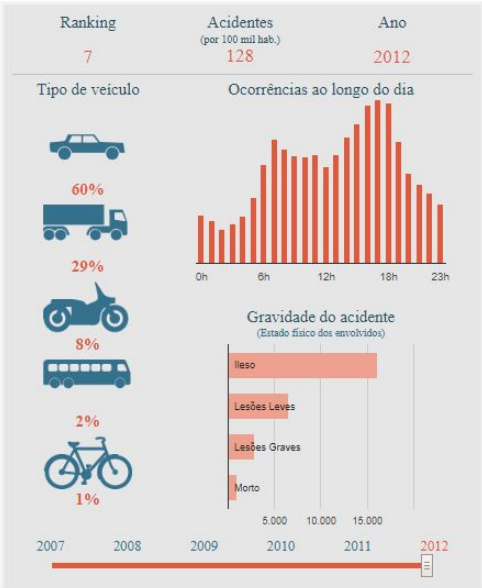
- ajuda o cidadão a monitorar a execução dos orçamentos municipal, estadual e federal
- acompanhamento e o entendimento de como estão sendo aplicados os recursos originados nos impostos e nas taxas que ele paga

Dataset	Período	Total de gastos	Número de Dados
Municipio de Curitiba	2015	R$ 25.509.621.547,39	188.673
Municipio de Recife	2015	R$ 4.122.505.297,25	102.992
Municipio de Recife	2016	R$ 2.743.731.903,77	45.205
Municipio de Curitiba	2016	R$ 5.935.067.757,55	28.022
Municipio de Belo Horizonte	2016	R$ 3.229.433.969,63	4.058
Municipio de Sao Paulo	2015	R$ 52.648.696.526,31	2.655
Municipio de Sao Paulo	2014	R$ 47.678.469.770,65	2.510
Municipio de Sao Paulo	2016	R$ 33.117.038.957,57	1.853
Municipio de Belo Horizonte	2015	R$ 9.128.986.036,51	929

Exemplos de Dados abertos

Ocorrências nas Rodovias Federais

- permite visualizar as ocorrências de acidentes nas Rodovias Federais



Ferramenta

Lemonade

- Uma plataforma para **criação visual e execução de fluxos de análise de dados.**

- Link: https://teste.ctweb.inweb.org.br/

O que posso fazer com o lemonade?

1. Criar um fluxo de processamento
    - Criar e customizar fluxos.
    - Estão disponíveis várias operações e filtros disponíveis para que você possa manipular seus dados.
2. Executar e gerenciar fluxos existentes
    - Todos os seus fluxos podem são gerenciáveis.
    - Pode criar, editar ou deletar.
    - Também é possível seguir o progresso dos fluxos que estão sendo executados, deletando ou pausando eles se quiser.

O que posso fazer com o lemonade?

3. Importar, exportar ou gerenciar datasets
    - Todos os seus dados podem ser importados em diferentes formatos para a aplicação, para serem manipulados.
    - Você pode gerenciar as bases de dados importadas e depois exportar o resultado de suas manipulações no formato que quiser
4. Visualização de dados
    - Estão disponíveis ferramentas para melhor visualizar os dados, abstraindo suas informações para gráficos ou outros métodos visuais

Exemplo 1 no Lemonade

Titanic

Titanic

- O naufrágio do Titanic é um dos naufrágios mais famosos da história.
- Essa tragédia **chocou a comunidade internacional** e levou a **melhores normas de segurança para os navios.**
- Uma das **razões** pelas quais o naufrágio levou a uma perda de vidas era que **não havia botes salva-vidas** suficientes para os passageiros e tripulantes.
- Alguns **grupos** de pessoas eram **mais propensos a sobreviver** do que outros: mulheres, crianças e a classe alta.
- Vamos tirar informações interessantes através do lemonade.

Dataset

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Flor	female	38	1	0	PC 17599	712.833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	84.583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	518.625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabet	female	27	0	2	347742	111.333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Ach	female	14	1	0	237736	300.708		C
11	1	3	Sandstrom, Miss. Marguerite Ru	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda A	female	14	0	0	350406	78.542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome	female	55	0	0	248706	16		S

?

Qual o gênero dos passageiros?

Gênero dos passageiros

Parâmetros da tarefa

Nome

Leitor de dados 1

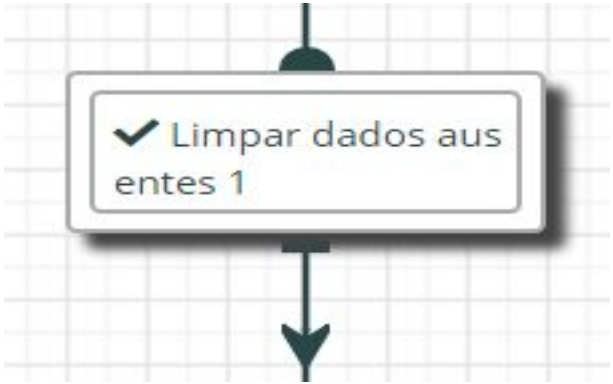
Execução

Fonte de dados*

Titanic

Tratar esses valores como nulos (separe por vírgula)

Inferir esquema da fonte de dados



Parâmetros da tarefa

Nome

Limpar dados ausentes 1

Execução

Atributo(s)*

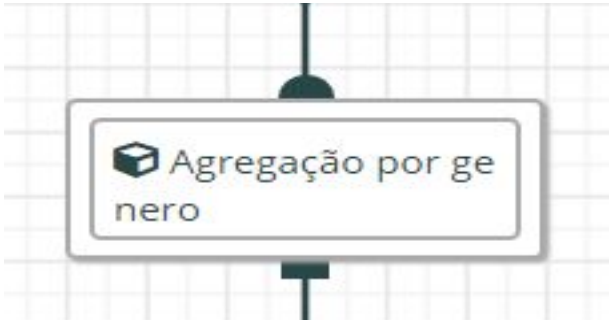
× class × Sex

Tipo de limpeza*

Remover toda a linha

Valor

Razão mínima de valores ausentes



Função de agregação

Função a ser aplicada aos dados agregados

Editor | Referências

Atributos	Função	Alias	
class	Count	class	

Adicionar

Ok

Parâmetros da tarefa

Nome

Agregação por genero

Execução

Selecione o(s) atributos para agregação *

× class

Função de agregação *

Abrir o Editor

Atributo usado como pivô

× sex

Valores do pivô (recomendado, se pivô for usado)

male,female

Atributo(s)

Selecione um ou mais atributos a serem usados como critério de ordenação. A ordem dos atributos faz diferença.

Editor | Referências

Atributos	Função	
class	Ascending	

Adicionar

Ok

Parâmetros da tarefa

Nome

Ordenar 1

Execução

Atributo(s) *

Abrir o Editor

Aparência

Comentário

Parâmetros da tarefa

Nome

Gráfico de barras 1

Execução

Atributo para eixo X*

× class

Atributos para eixo Y (cada um será uma série) *

× male × female

Título

Genero por Classe

Completo

Código | Workflow | Relatório

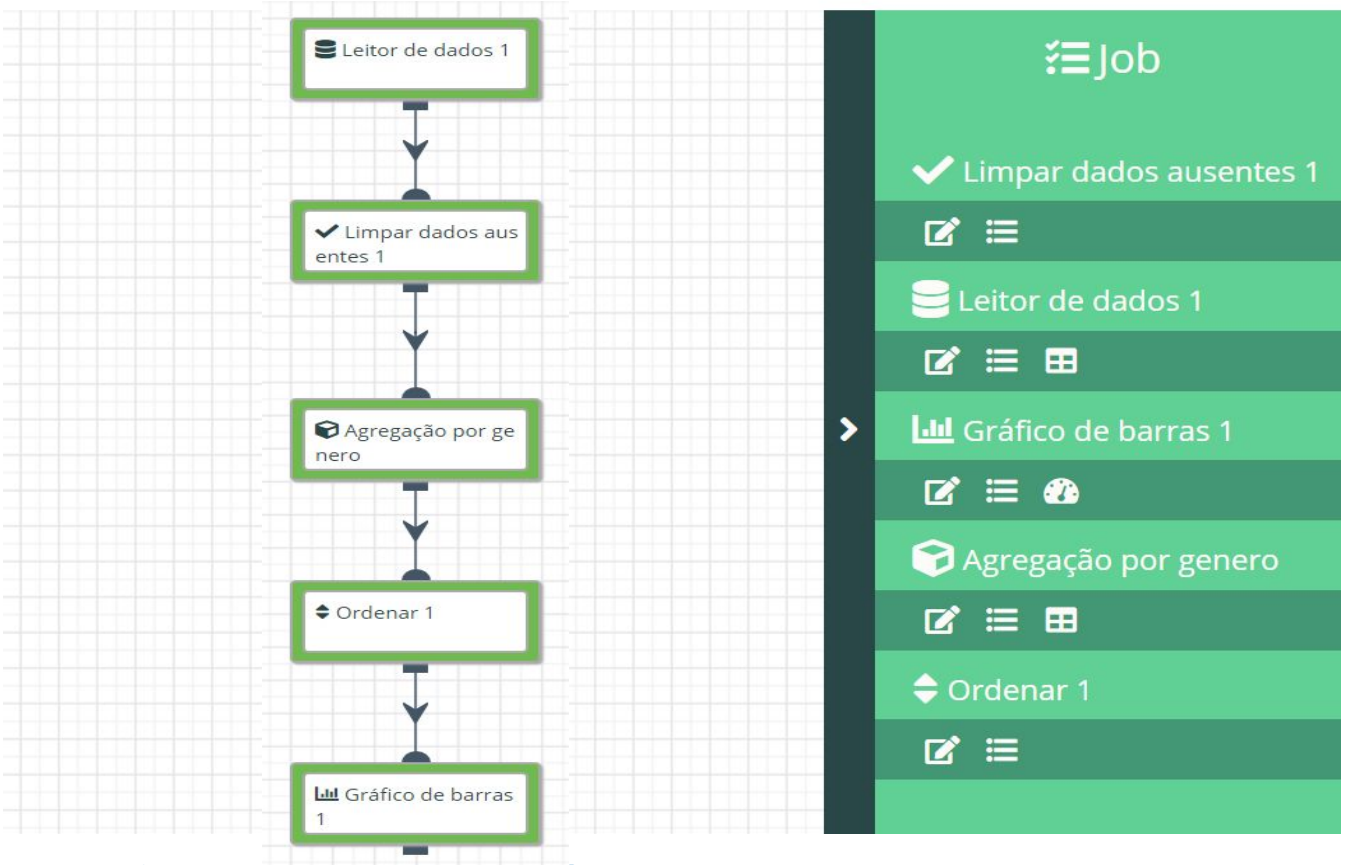
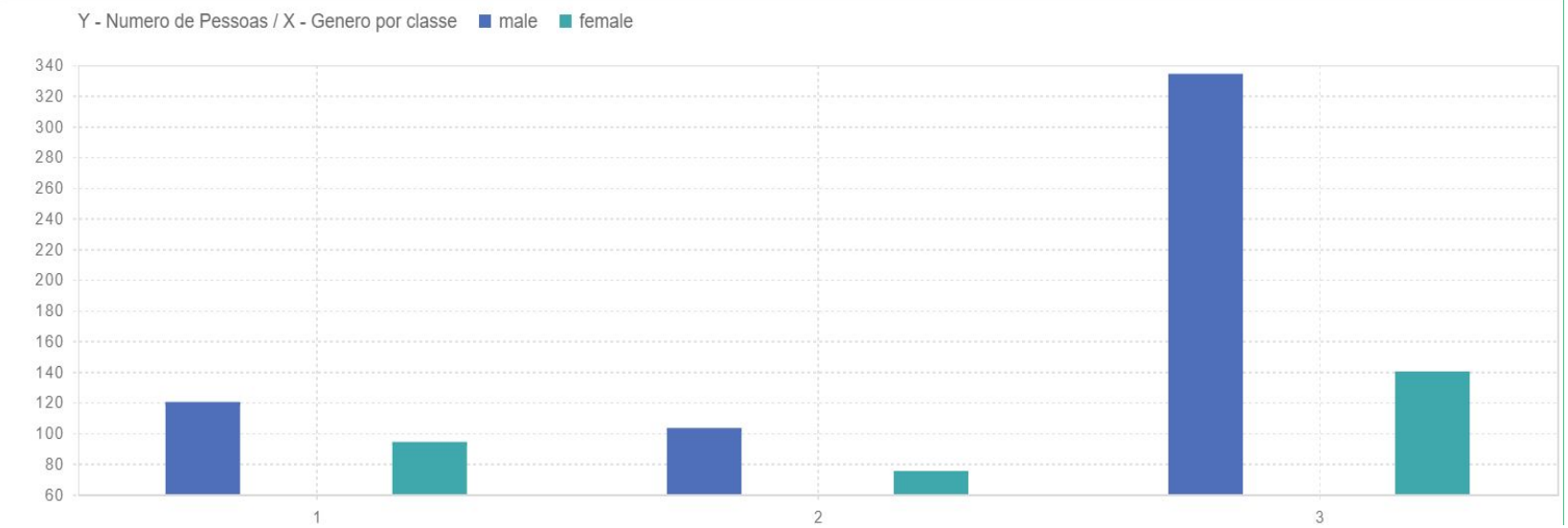


Gráfico de barras 1 Completo

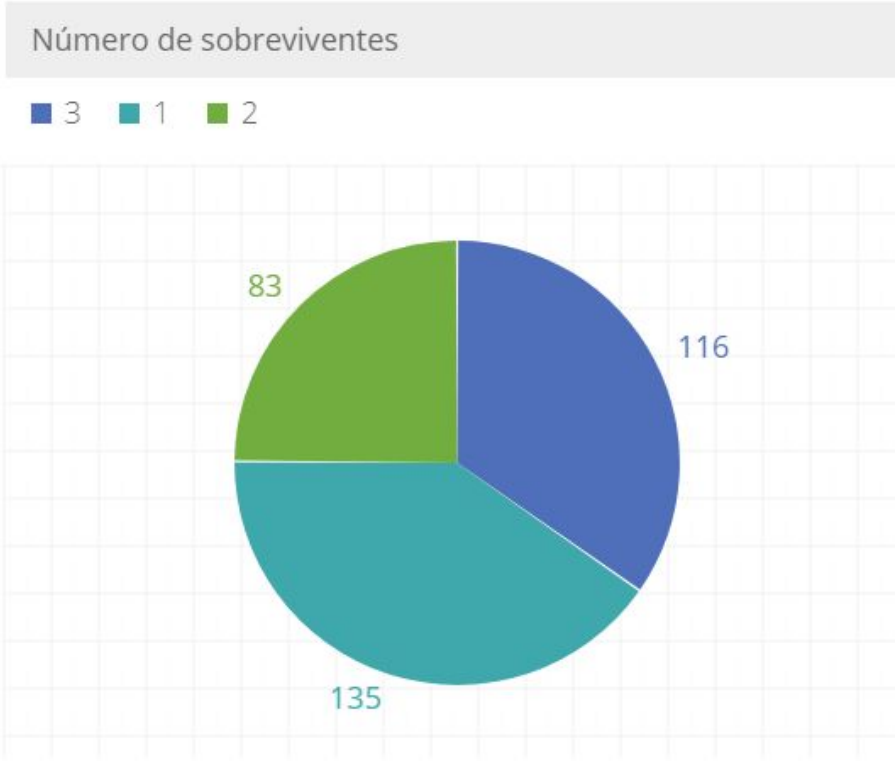
Resultado Logs Parâmetros

Genero por Classe



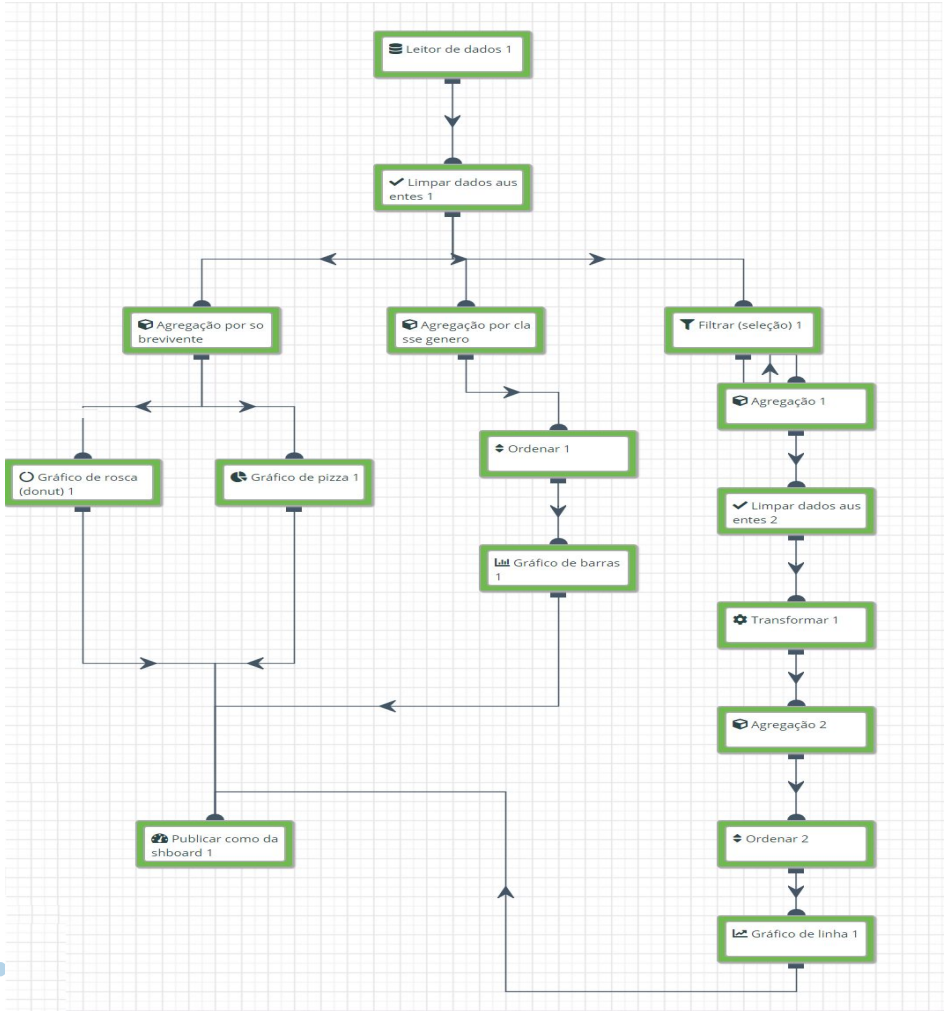
Quem são os sobreviventes e os mortos?

Agregação por Sobreviventes



Fluxo completo

https://teste.ctweb.inweb.org.br/home/workflows/556/draw



Exemplo 2 no Lemonade

Dengue - Região

Dados Ano 2013

?

Quais regiões do Brasil a dengue tem maior incidência?

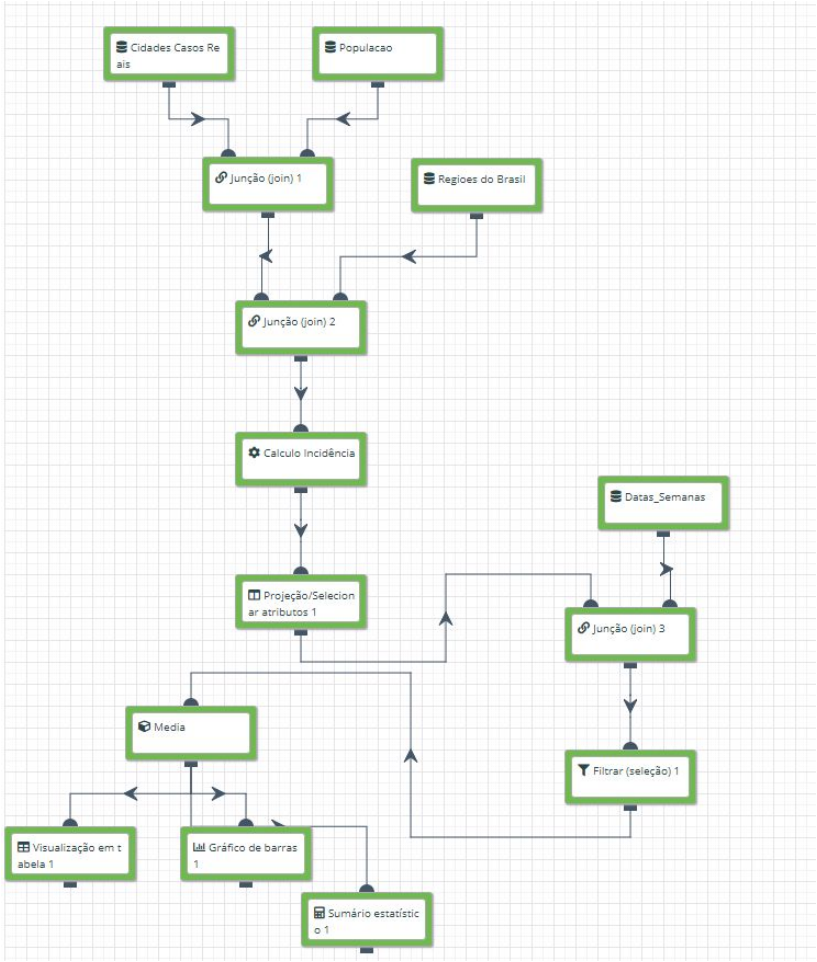
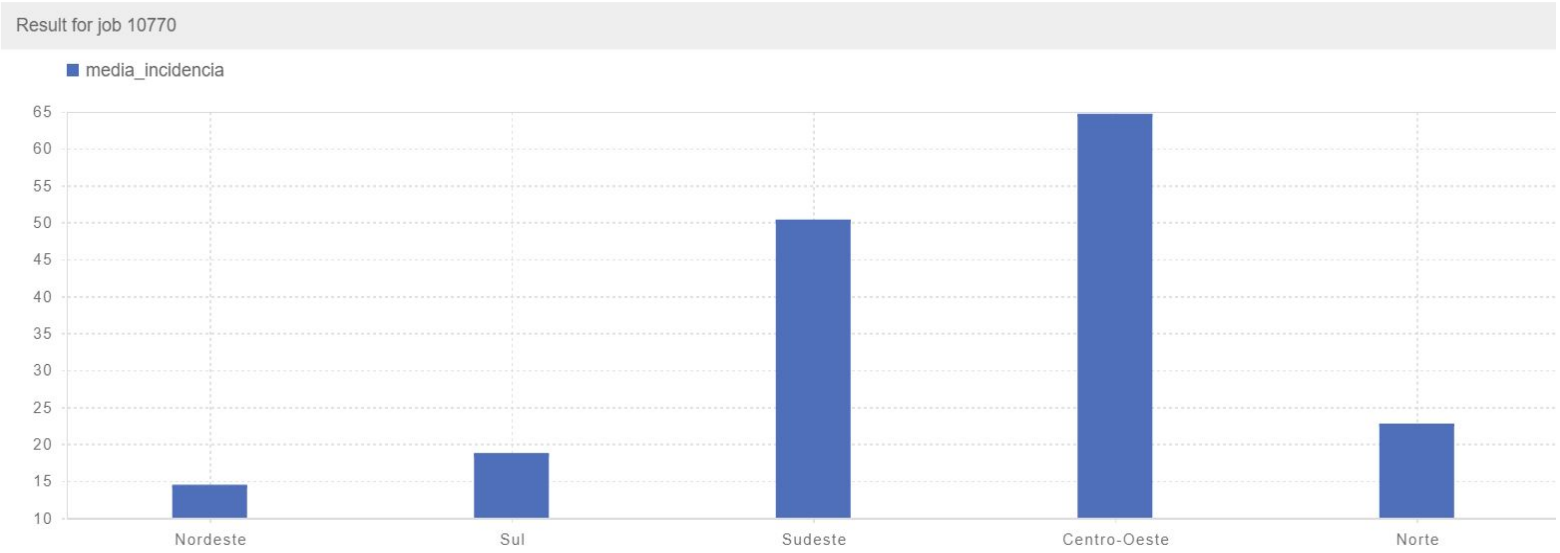

Cidades Casos Reais
Populacao
Junção (join) 1
Regioes do Brasil
Junção (join) 2
Calculo Incidência
Datas_Semanas
Projeção/Selecionar atributos 1
Junção (join) 3
Media
Filtrar (seleção) 1
Visualização em tabela 1
Gráfico de barras 1
Sumário estatístico 1

Gráfico de barras 1 ✔ Completo

As regiões Centro-Oeste e Sudeste lideram em número notificações.

Resultado | Logs | Parâmetros



https://www.paho.org/bra/index.php?option=com_content&view=article&id=3159:dados-da-dengue-no-brasil-2013&Itemid=463

Exemplo 3 no Lemonade

Dengue - Clima

Dados Ano 2013

?

Quais os períodos de maior incidência de dengue no país?

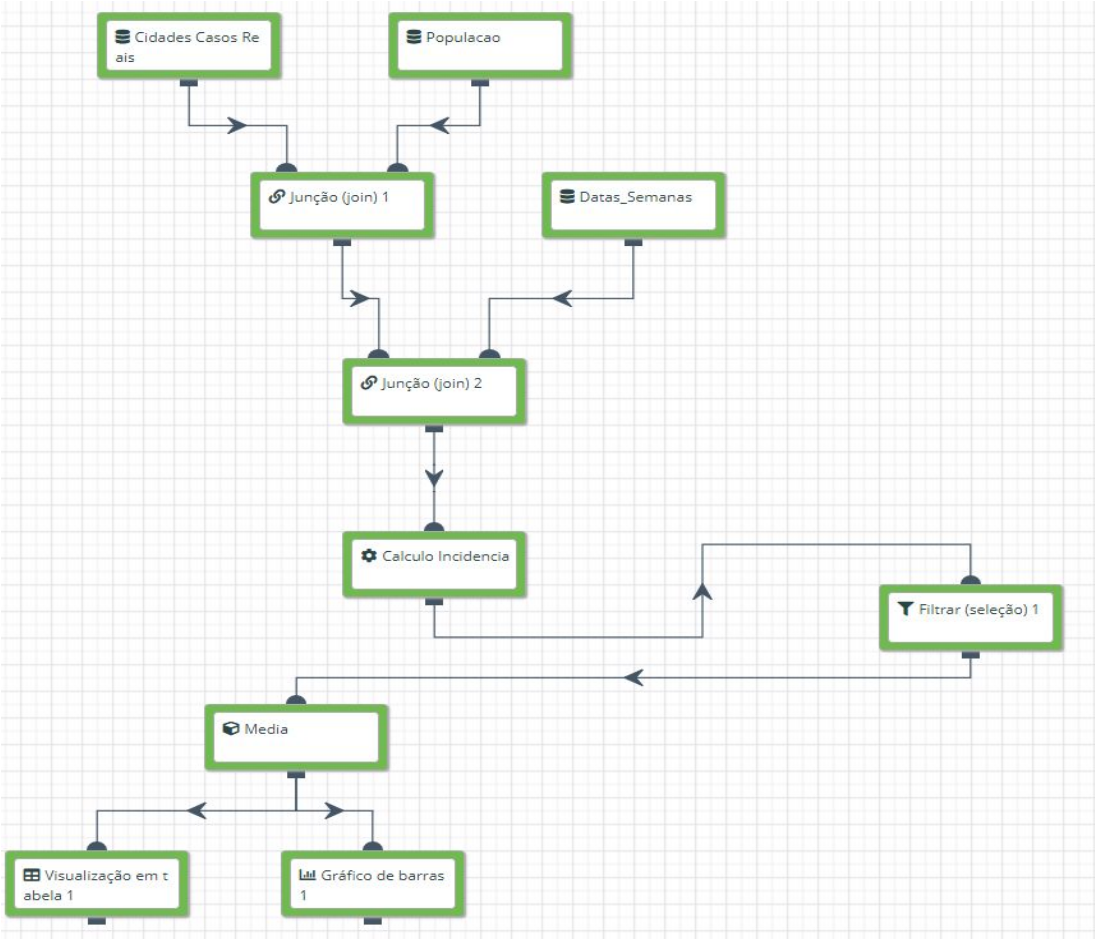
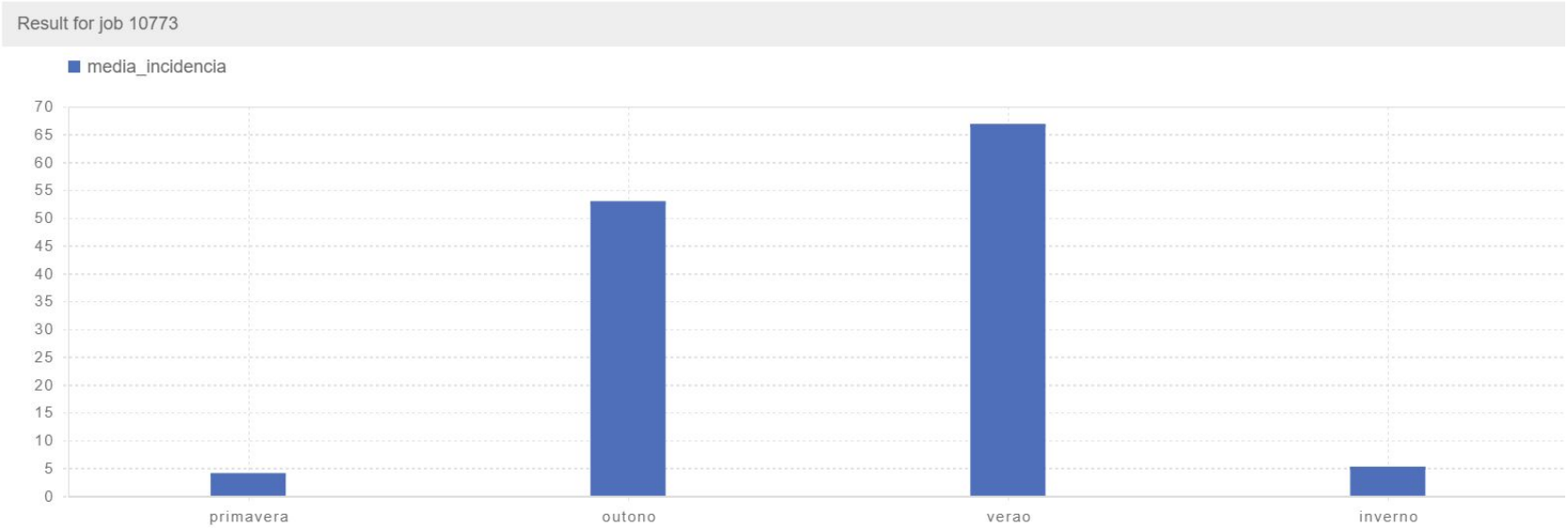

Cidades Casos Reais
Populacao
Junção (join) 1
Datas_Semanas
Junção (join) 2
Calculo Incidencia
Filtrar (seleção) 1
Media
Visualização em tabela 1
Gráfico de barras 1

Gráfico de barras 1 ✔ Completo

Chegada do verão aumenta o sinal de alerta contra a dengue

Resultado | Logs | Parâmetros



http://www.gazetaonline.com.br/cbn_vitoria/reportagens/2017/12/chegada-do-verao-aumenta-o-sinal-de-alerta-contra-a-dengue-1014110428.html

Cientista de dados

- Matemática e Estatística
- Programação e Banco de dados
- Comunicação e Visualização
- Conhecimento de Negócios
- Gostar de resolver problemas
- Ser curiosa
- Ter pensamento lógico
- Ser estratégica, proativa, criativa, inovadora
- e colaboradora

Motivo de ser considerada a profissão do futuro

- “Big data analytics é o próximo mercado de 1 trilhão de dólares!” Michael Dell
- Chamada de a mais sexy deste século pela Harvard Business Review
- Existem muitos dados que ainda precisam ser coletados e analisados
- As empresas dependem cada vez mais da análise de dados para impulsionar a tomada de decisões
- Falta profissionais qualificados

Motivo de ser considerada a profissão do futuro

- Em média, um valor bruto de 10 mil reais, mas algumas empresas brasileiras chegam a pagar 25 mil para os funcionários da área (Nubank)
- Um estudo realizado pela International Business Machines (IBM) apontou que a demanda por empregos na área de data science deve aumentar em mais de 15% ao redor do mundo, e, aparentemente, o Brasil vem acompanhando essa tendência.

Perguntas?

Obrigada!

Bárbara Silveira Fraga - barbarasilveira@dcc.ufmg.br
Karen Martins - karensm@dcc.ufmg.br
Larissa Maia - larissaemanuelle@dcc.ufmg.br