



DATA DISCOVERY, OLAP E VISUALIZAÇÃO DE DADOS

Marcelo Tavares de Lima

Data Discovery, Olap e visualização de dados

1^a edição

Londrina
Editora e Distribuidora Educacional S.A.
2019

© 2019 por Editora e Distribuidora Educacional S.A.

Todos os direitos reservados. Nenhuma parte desta publicação poderá ser reproduzida ou transmitida de qualquer modo ou por qualquer outro meio, eletrônico ou mecânico, incluindo fotocópia, gravação ou qualquer outro tipo de sistema de armazenamento e transmissão de informação, sem prévia autorização, por escrito, da Editora e Distribuidora Educacional S.A.

Presidente

Rodrigo Galindo

Vice-Presidente de Pós-Graduação e Educação Continuada

Paulo de Tarso Pires de Moraes

Conselho Acadêmico

Carlos Roberto Pagani Junior

Camila Braga de Oliveira Higa

Carolina Yaly

Giani Vendramel de Oliveira

Juliana Caramigo Gennarini

Nirse Ruscheinsky Breternitz

Priscila Pereira Silva

Tayra Carolina Nascimento Aleixo

Coordenador

Nirse Ruscheinsky Breternitz

Revisor

Eduardo Mayer Fagundes

Editorial

Alessandra Cristina Fahl

Beatriz Meloni Montefusco

Daniella Fernandes Haruze Manta

Hâmila Samai Franco dos Santos

Mariana de Campos Barroso

Paola Andressa Machado Leal

Dados Internacionais de Catalogação na Publicação (CIP)

L99d Lima, Marcelo Tavares de
 Data Discovery, Olap e visualização de dados/ Marcelo
 Tavares de Lima, – Londrina: Editora e Distribuidora
 Educacional S.A. 2019.
 133 p.

ISBN 978-85-522-1582-0

1. Dashboard Design. 2. Big Data. I. Lima, Marcelo
Tavares de. II. Título.

CDD 004

Thamiris Mantovani CRB: 8/9491

2019

Editora e Distribuidora Educacional S.A.
Avenida Paris, 675 – Parque Residencial João Piza
CEP: 86041-100 — Londrina — PR
e-mail: editora.educacional@kroton.com.br
Homepage: <http://www.kroton.com.br/>



SUMÁRIO

Apresentação da disciplina	05
Os Principais métodos de Visualização de Dados	06
A organização visual (Visualização de dados e <i>big data analytics</i>)	23
O processo de Design de <i>Dashboard</i>	42
Visualização de dados com R, Python e Qlik Sense	62
Visualização de dados utilizando ferramentas OLAP	84
<i>Data discovery</i>	102
Outras ferramentas para visualização de dados (Chart.js, Leaflet, Datawrapper, Dygraphs, Highcharts, Google Charts, Polymaps e Weka)	120



► Apresentação da disciplina

Seja bem-vindo ao mundo da visualização de dados e de *Big Data*!

Esta disciplina apresentará a você os principais conceitos e definições de visualização de dados, indicando os parâmetros importantes para realizar uma escolha apropriada para transmitir um resultado a um público alvo, através de gráficos e outros recursos visuais. A intenção do uso desse recurso é facilitar a apresentação de resultados importantes obtidos em pesquisas e em empresas, como o atingimento de metas, dentre outros.

Também serão apresentadas algumas ferramentas, dentre muitas existentes, para a elaboração de imagens que apresentarão suas informações de maneira clara, sucinta e eficiente. Dentre as ferramentas a serem apresentadas, algumas exigirão um certo conhecimento de linguagem de programação, como JavaScript, linguagem R e linguagem Python. Enquanto que outras ferramentas não possuem essa exigência.

Serão apresentados, também, alguns conceitos de arquiteturas de dados como as ferramentas OLAP, apropriadas para consulta de grandes bases de dados, através de recursos diversos a depender da estrutura física disponível para a sua realização.

Desejamos que você aproveite bastante este momento de estudo do conteúdo. Que ele possa trazer *insights* para a sua vida, tanto acadêmica quanto profissional, e que ao final deste curso você possa sair com um diferencial em sua formação!

Bons estudos!



Os principais métodos de visualização de dados

Autor: Marcelo Tavares de Lima

► Objetivos

- Descrever um breve histórico e a importância da visualização de dados.
- Apresentar os principais métodos de visualização de dados.
- Apresentar aplicações diversas dos diferentes métodos de visualização de dados.

1. Introdução

Neste texto serão apresentados os principais métodos de visualização de dados utilizados em negócios e, também, no meio acadêmico. A intenção é fazer um aparato geral sobre os métodos que serão apresentados e descrever aplicações para exemplificar.

“Muitos afirmam que a informação é a nova moeda dos negócios, e que a internet é a agência de câmbio na qual ela é negociada” (UNDERS SERVERS E DATACENTERS, 2018, n.p.). Uma das principais ferramentas para a divulgação de informação é a informação gráfica, ou a visualização de dados. O motivo para a crescente valorização do seu uso é diverso, no entanto, é totalmente compreensível dado o contexto em que vem sendo utilizada. Você verá ao longo deste texto!

Estamos na era dos dados! Mais do que nunca se criam inovações e muito mais oportunidades para trabalhar com dados nunca vistos antes na história da humanidade. A transformação de resultados analíticos em imagens é uma consequência do avanço das tecnologias e dos métodos de análise.

“Um dos pioneiros de visualização de dados, William Cleveland, escreveu que as visualizações traduzem números em imagens, seja em papel ou em tela, tal que o leitor possa traduzi-la em seu cérebro” (GRANT, 2019, p. 21, tradução nossa).

Ao final deste texto, você estará familiarizado com os principais métodos de visualização de dados, o que possibilitará a você escolher aquele mais apropriado para os seus interesses e objetivos, tanto profissionais quanto acadêmicos.

2. Métodos de visualização de dados

É sabido que os dados estão em todos os lugares. Desde muito tempo, os dados produzidos são armazenados e processados para que produzam informações que possam subsidiar decisões. No entanto, mais do que nunca, dados estão sendo produzidos em larga escala. A corrida para o armazenamento e o tratamento de dados, no intuito de torná-los em informações úteis, está bastante acelerada.

Os métodos de visualização de dados têm acompanhado essa corrida desde o século XV, juntamente com os métodos analíticos e algébricos, no processo evolutivo tecnológico, tornando-se cada vez mais frequentes nos relatórios finais de análise de dados.

Algumas características dos dados e resultados só podem ser identificadas através de uma representação gráfica, ou, pelo menos, podem ser identificadas com maior facilidade. O poder que a visualização de dados possui está na capacidade de transformação de dados brutos em algo visual para trazer um significado relevante para o entendimento de algo que se pesquisa, além de facilitar a compreensão da mensagem que se deseja transmitir com o resultado obtido.

As empresas e/ou pesquisadores, para obterem um aprendizado com os seus dados, necessitam adotar a visualização como ferramenta para a exploração e para a comunicação entre os interessados. O uso dessa metodologia ocorre, principalmente, quando se supõe que existam resultados que apresentam algo de interessante e significativo após o seu tratamento analítico.

Com o avanço dos recursos tecnológicos, os métodos de exploração de dados visualização e gráficos também evoluíram. Evoluíram, também, os métodos de visualização ou gráficos. Tem-se, a cada dia, mais recursos disponíveis, que vão desde recursos pagos, gratuitos, recursos *online*, de código aberto, dentre outros.

Existe uma disputa acirrada para o armazenamento de grandes volumes de dados, conhecidos como *big data*, e não só é disputado o armazenamento eficiente, mas, também, o tratamento e a divulgação adequada para este tipo de informação.

Dentro do contexto de *big data* é importante que não se perca ou não fique sem saber o que fazer com o grande volume de dados gerados, pois não basta ter espaço para armazenamento, é necessário, também, que se saiba utilizar um tratamento analítico e gráfico apropriados para divulgação e extração de informações importantes.

A análise gráfica de grandes volumes de dados, com o uso da ferramenta adequada, ajuda na identificação de padrões e tendências. É por isso que a apresentação visual ou gráfica se torna um importante aliado, já que muitos especialistas da área afirmam que essa é a melhor maneira de garantir que os resultados alcancem o maior público possível.

Por muito tempo, planilhas de dados, como a planilha Microsoft Excel, por exemplo, foram as ferramentas mais utilizadas para o tratamento de dados e produção de informações visuais, dada a facilidade de sua manipulação. Essas ferramentas ainda são bastante utilizadas, no entanto, em se tratando de *big data*, ferramentas com maior capacidade, tanto de armazenamento quanto analítica, têm sido desenvolvidas. Apesar disto, a planilha Microsoft Excel ainda é listada como uma das ferramentas de *business analytics*.

O armazenamento de dados através de bancos de dados tornou-se cada vez mais rotineiro, de modo que muitas empresas passaram a investir em ferramentas tradicionais de *Business Intelligence* (BI), as quais garantiam possuir a capacidade de ampla integração, análise e apresentação de resultados do tratamento de dados.

Inicialmente, os produtos das ferramentas de BI eram relatórios e, em seguida, os chamados *dashboards*, os quais passaram a ser cada vez mais interativos com o avanço do desenvolvimento do BI.

Em tempos de *big data*, onde a exigência por ferramentas com maior capacidade analítica e de apresentação visual é cada vez maior, as ferramentas tradicionais de BI deixam de atender de maneira eficiente as necessidades, principalmente, as necessidades das grandes empresas, que produzem grandes volumes de dados.

Com a intenção de aumentar o poder analítico da manipulação e tratamento de grandes volumes de dados, o mercado de visualização cresceu bastante e vem em uma tendência crescente já há um certo tempo, aumentando o investimento em processos, pessoas e em tecnologia.

A visualização de dados tem a intenção de facilitar a transmissão de um resultado encontrado com o tratamento aplicado a um conjunto de dados. Por isso, um gráfico, mapa ou qualquer outra imagem, precisa transmitir com clareza, facilidade e rapidez a sua mensagem, sem deixar o leitor e visualizador confuso, senão não atinge o seu objetivo geral, que é a compreensão e a comunicação. Dentre os tipos gráficos mais utilizados estão os gráficos de barras verticais e barras horizontais, linhas, setores, diagrama de dispersão, etc. O tipo de visualização a ser utilizada depende diretamente do tipo de informação que está sendo utilizado para a sua elaboração.

PARA SABER MAIS



Como ler um *scatter plot* ou diagrama de dispersão: qualquer ponto em um *scatter plot* pode ser localizado observando o quanto longe ele se encontra dos eixos principais de um plano cartesiano (eixo-x e eixo-y). Para

isso, é necessário que existam marcas de escala nos dois eixos. Apresentar o gráfico com linhas de grade pode facilitar a identificar a localização com maior precisão de cada ponto, o qual representa um par ordenado com dois valores de duas variáveis quantitativas. Ele é apropriado para mostrar a associação entre essas medidas.

O termo ‘visualização de dados’ (*data visualization* ou *dataviz*) atrai atenção de jornalistas, acadêmicos e pessoas de negócios, de maneira igual, pois é um método que pode ser utilizado em qualquer meio. Por que uma imagem pode ser tão poderosa quando comparada a números? (GRANT, 2019, p. 3, tradução nossa)

Essa é uma questão que deve sempre estar na mente daqueles que fazem ou pretendem fazer uso de ferramentas de visualização de dados. Pensar desta maneira ajuda a valorizar e a motivar na produção de uma boa imagem.

Basicamente, o produto de uma elaboração visual de um conjunto de dados é um gráfico. É claro que, em tempos de *big data*, a elaboração de um gráfico tem disponível mais recursos tecnológicos, os quais disponibilizam mais possibilidades e muito mais ferramentas para a sua produção. No entanto, para uma correta utilização, de qualquer ferramenta que seja de elaboração de um resultado visual ou gráfico, é preciso reconhecer o tipo de informação que está sendo manipulada.

Os dados precisam ser caracterizados corretamente, como, por exemplo, em relação à dimensão em que foram tratados e a natureza do domínio no qual estão definidos. O domínio de um conjunto de dados pode ser, segundo Freitas et al. (2001), contínuo, contínuo-discretizado ou discreto, e quanto à dimensão, segundo os mesmos

autores, pode ser unidimensional, bidimensional, tridimensional ou n-dimensional. O Quadro 1 resume os principais critérios de caracterização de conjuntos de dados que precisam ser conhecidos para uma escolha correta e apropriada para elaboração de uma visualização.

Quadro 1 – Sumário da caracterização de dados

Critério	Classe	Exemplo
Classe de informação	Categoría	Gênero ou sexo
	Escalar	Temperatura
	Vetorial	Grandezas físicas associadas a dinâmica de fluidos
	Relacionamento	Link num hiperdocumento
Tipo de valores	Alfanumérico	Gênero ou sexo
	Numérico	Temperatura
	Simbólico	Link num hiperdocumento
Domínio	Discreto ou categórico	Marca de automóveis
	Contínuo	Superfície de um terreno
	Contínuo-discretizado	Anos (tempo discretizado)
Dimensão	1D	Fenômeno ocorrendo no tempo
	2D	Superfície de um terreno
	3D	Volume de dados médicos
	n-D	Dados de uma população

Fonte: adaptado de Freitas et al. (2001).

As ferramentas analíticas para produzir informações visuais ou gráficos disponíveis na era “big data” permitem a elaboração de gráficos mais complexos. No entanto, assim como há a necessidade de conhecer a dimensão e o domínio da informação, também se faz necessário reconhecer o tipo de representação gráfica apropriada para cada tipo, e o Quadro 2 apresenta um resumo sobre isso.

Quadro 2 – Classes de representações visuais

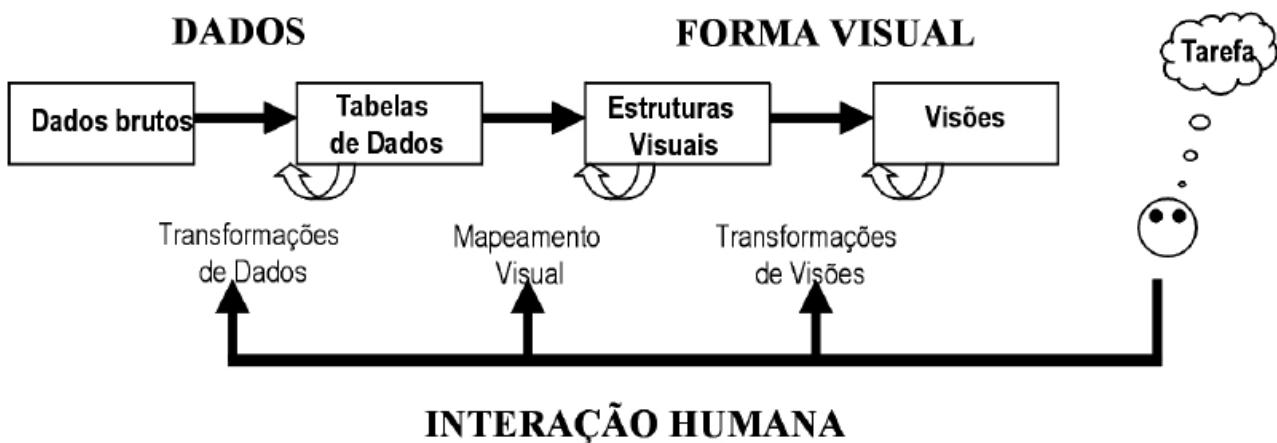
Classe	Tipo	Utilização
Gráficos 2D e 3D	Pontos circulares	Representação da distribuição dos elementos no espaço domínio, representação da dependência/correlação entre atributos.
	Linhas	
	Barras	
	Superfícies (para 3D)	
Ícones Glifos Objetos geométricos	Elementos geométricos 2D ou 3D diversos	Representação de entidades num contexto, representação de grupos de atributos de diversos tipos.
Mapas	De pseudo-cores	Representação de campos escalares ou de categorias.
	De linhas	Representação de linhas de contorno de regiões, isovalores.
	De superfícies	Representação de entidades num contexto, representação de grupos de atributos de diversos tipos, no espaço 3D.
	De ícones, símbolos diversos	Representação de grupos de atributos (categorias, escalares, vetoriais, tensoriais).
Diagramas	Nodos e arestas	Representação de relacionamentos diversos: É-um, É-parte-de, Comunicação, Sequência, Referência, etc.

Fonte: adaptado de Freitas et al. (2001).

Para se obter uma visualização de dados bem elaborada e bem-sucedida nos seus objetivos, além de se ter o cuidado com as características apresentadas nos Quadros 1 e 2; também é necessário fazer uso adequado de propriedades como tamanho, cor e forma gráfica adequada para tipo de informação tratada e que se deseja apresentar. É necessário, também, levar em conta a história que se deseja passar, o que se deseja enfatizar visualmente e a quem se deseja atingir, ou seja, o público-alvo.

O ciclo de tarefas necessárias para se chegar a um produto gráfico de visualização de dados pode ser generalizado conforme apresentado em Freitas et al. (2001), os quais o chamam de modelo de Card, apresentado na Figura 1.

Figura 1 – Modelo de referência de visualização de Card



Fonte: Freitas et al. (2001).

Ao longo de cada uma das etapas do modelo de Card é preciso ter muito claro o tipo de informação que está em tratamento para que se possa utilizar a ferramenta mais apropriada para se obter um produto de qualidade e livre de vieses e erros. No modelo de Card é visível que a interação humana ocorre ao longo de todo o processo de elaboração do produto final, a visualização. Portanto, é extremamente importante ter pessoal qualificado e habilitado a manusear bem todas as ferramentas e a ter o conhecimento teórico necessário. Como, por exemplo, as habilidades podem acontecer a partir de cursos de graduação como ciência da computação, estatística, cursos de complementação como *data mining*, *design* gráfico e, enfim, a interação necessária entre homem e máquina.

“O processo de criação de uma visualização de dados ou um infográfico é multidisciplinar e inclui uma grande variedade de subprocessos, que devem ser bem integrados para que se tenha êxito” (ACCENTURE, 2014, p.6). Os subprocessos citados podem ser apresentados de forma resumida e esquemática, conforme o Quadro 3.

Quadro 3 – Etapas do processo para elaboração de visualização de dados

Processo	Habilidade necessária	Resumo
Definição de objetivo		Entender a motivação e definir um objetivo.
Adquirir	Ciência da computação	Adquirir dados relevantes e os mais completos possível. Caso necessário, completar com dados públicos.
Formatar	Ciência da computação	Analizar e formatar os dados obtidos em algum formato adequado para o uso. Caso sejam vários conjuntos de dados, garantir a integração entre eles.
Filtrar	Estatística e data mining	Filtrar os dados, para que o conjunto de dados contenha apenas o que se deseja trabalhar.
Analisar	Estatística e data mining	Escolha de ferramenta apropriada para análise. Em seguida, modelar e analisar os dados. Elaborar visualização exploratória e, se for o caso, reiterar as etapas anteriores.
Representar	Designer gráfico	Escolha de ferramenta apropriada para a visualização dos dados e elaborar o infográfico.
Refinar	Designer gráfico	Refinar a visualização dos dados ou infográfico para adequação ao público-alvo.
Interagir	Interação homem-máquina	Publicar, implantar e interagir com a visualização de dados. Se o produto final for insatisfatório, determinar a etapa do processo à qual retomar e, então, repeti-las. Caso contrário, encerrar o processo.

Fonte: Accenture (2014, p. 6).

Sabe-se que em tempos de “*big data*” as coisas mudam muito rapidamente. No entanto, é importante citar alguns programas computacionais que podem ajudar a desbravar o tratamento de informações para transformá-las em imagens. Grant (2019, p. 15, tradução nossa) afirma que “se você deseja se tornar um elaborador versátil de visualizações, recomendo tornar-se familiarizado com algumas ferramentas computacionais”.

Dentre as ferramentas citadas por Grant (2019), pode-se, além de citar os nomes, indicar qual a funcionalidade da ferramenta. Por exemplo,

para tratamento estatístico de dados, ele cita programas como o Stata, as linguagens R e Python. Para visualização rápida de dados, cita o Tableau. Para edição de arquivo SVG, cita Inkscape e Illustrator, dentre outros.

Além dos programas citados por Grant (2019), também vale a pena ter conhecimento sobre: Infogram, Plotly, RAW, D3.js, Google charts, Qlick Sense, etc. Sobre cada um destes programas será apresentada uma descrição de suas utilidades.

ASSIMILE



Esteja preparado para rascunhar e descartar! A escolha pela melhor forma de visualização nasce dos ensaios e rascunhos que se faz até escolher aquela que melhor poderá representar a informação que se deseja transmitir e o público-alvo desejado.

Apesar de terem sido citados por Grant (2019) como programas estatísticos, o Stata, o R e o Python também são ferramentas produtoras de gráficos e imagens. No entanto, o seu uso requer um mínimo de conhecimento de linguagem de programação. Se bem que já existem versões para o Stata e interfaces gráficas para o R que superam a necessidade dessa exigência, podendo ser utilizados através de menu de opções.

Tableau: Atualmente é a ferramenta mais popular, segundo alguns especialistas. Sua popularidade se deve por suportar ampla variedade de gráficos, mapas, tabelas e outros elementos. Outro motivo da sua popularidade é a não exigência de conhecimento de linguagem de programação. Possui versões gratuitas e pagas e que se diferenciam segundo algumas funcionalidades e ferramentas.

Infogram: Apresenta fácil utilização para produção de mapas interativos, gráficos e infográficos. Infográficos são subconjuntos de visualização de dados, ou seja, são imagens elaboradas com dados específicos e divulgados, em geral, em jornais e revistas. Possui versão gratuita, no entanto, esta apresenta muitas restrições quanto à disponibilidade de ferramentas para elaboração de imagens. Também não exige conhecimento sobre linguagem de programação.

Plotly: É uma ferramenta elaborada para analisar e visualizar dados na web. Oferece uma ampla variedade de gráficos com recursos para compartilhamento social. A estética de suas imagens é considerada uma das mais profissionais do mercado. Existe versão gratuita e paga da ferramenta. Para ser manipulado, exige um certo grau de conhecimento de linguagem de programação Python.

RAW: Ferramenta para elaboração de mapas e diagramas visuais com o uso de Google Docs, planilhas Microsoft Excel e similares. Construído sobre biblioteca interativa em JavaScript D3.js, possui interface bastante interativa, o que facilita o seu uso. É processado no navegador de internet. Permite o auxílio para melhoria de gráficos das ferramentas Illustrator e Inkscape. É uma ferramenta totalmente gratuita.

Google Charts: Permite a criação de gráficos e infográficos. Possui interface bastante amigável e uma ampla galeria de modelos e configurações disponíveis.

Qlick Sense: De certa forma se assemelha a Tableau, pois possui ferramentas semelhantes. Possui versões com funcionalidades distintas e pagas, como QlikView, Qlik Insight Bot, etc. Também existe versão de teste, que é gratuita por determinado período.

Apesar de o Tableau ser considerado por muitos especialistas em visualização de dados como a principal ferramenta para a produção de visualização de dados, no entanto não existe uma ferramenta ideal, pois

a ideal é aquela que atende aos objetivos específicos de um trabalho em execução.

A importância de conhecer mais de uma ferramenta ocorre porque não existe uma que atenda a todas as necessidades de todos os trabalhos possíveis. Para isso, quando uma ferramenta não atender algo, deve-se buscar em outras o que se deseja executar.

Publicação da Accenture (2014) apresenta uma divisão para as ferramentas de visualização de dados: ferramentas de *business intelligence* (BI), ferramentas analíticas, ferramentas de visualização e ferramentas para trabalho personalizado. As principais diferenças entre elas e os programas computacionais apropriados para cada tipo são apresentados a seguir:

Ferramentas de BI: São as ferramentas mais utilizadas para a visualização de dados. São apropriadas para elaboração de relatórios e, de certa forma, para elaboração de *dashboards*. Por tender a ter painéis estáticos ou limitados, funciona bem para apresentar tabelas e gráficos mais simples e padronizados. São consideradas ineficientes para a etapa de análise exploratória de dados e para aplicações interativas. Há um movimento entre as empresas de BI para superar essas limitações. As ferramentas que compõem esse grupo são alguns produtos da Microsoft e IBM, softwares como o SAS, SAP e MicroStrategy.

Ferramentas analíticas: São apropriadas para executar análises estatísticas de dados. No entanto, possuem capacidade limitada de visualização de dados. No geral, são ferramentas que exigem conhecimento de programação. São mais utilizadas para testar modelos e, menos para apresentação visual de dados. As ferramentas que compõem esse grupo são R, SPSS, SAS, Statistica, Minitab e Matlab.

Ferramentas de visualização: Possuem funcionalidades avançadas de visualização de dados, mas capacidade analítica menos sofisticada. São

mais apropriadas para a etapa exploratória simples. As ferramentas que compõem esse grupo são Tableau, Spotfire, QlickView e Advizor.

Ferramentas para trabalho personalizado: Exigem habilidades avançadas e conhecimento técnico específico para serem implementadas. São utilizadas para apresentação de tema específico ou para trabalho pontual. As ferramentas que compõem esse grupo são D3, Processing e Adobe Illustrator.

Dado o exposto, é possível perceber que há uma infinidade de recursos tecnológicos para a produção de visualização. O que está apresentado nesta leitura fundamental é pouco diante do que existe. Este texto apresentou um contexto sobre métodos de visualização de dados, as principais ferramentas existentes e as diferenças entre elas. Apresentou, também, as características importantes a serem consideradas na etapa de elaboração de informações visuais, os chamados infográficos.

TEORIA EM PRÁTICA

Muitos meios de comunicação divulgam notícias sobre as empresas que utilizam métodos de *big data* e, também, sobre empresas que oferecem o serviço de *big data*.

Para exemplificar, uma matéria do jornal Estadão, na seção de economia e negócios divulgada em 20/05/2018, apresenta *startups* que oferecem serviços de *big data*, os quais incluem coleta, tratamento, visualização e enriquecimento de dados (ESTADÃO, 2018).

Em época de produção de grandes volumes de dados, muitas empresas passaram a terceirizar a organização da massa de dados que produzem, dando espaço para *startups* como a Deep Center, especializada em gestão de dados para escritórios de cobrança e *contact centers*.

Uma das ferramentas oferecidas pela empresa é o Data Discovery, plataforma de visualização, exploração e análise de dados. Não é a única ferramenta disponível em sua gama de serviços de *big data*. O portal da empresa apresenta vários outros.

VERIFICAÇÃO DE LEITURA



1. Grant (2019) declarou que um determinado especialista em visualização de dados afirmou que as visualizações traduzem os números em imagens. Qual o nome deste profissional? Assinale a alternativa correta.
 - a. William Crew.
 - b. Robert Grant.
 - c. William Cleveland.
 - d. Robert Redford.
 - e. Joseph Tableau.

2. O tratamento analítico gráfico dispensado a grandes volumes de dados ou “*big data*”, quando realizado de forma correta e com a ferramenta apropriada, fornece vários benefícios como resultado. No entanto, um aspecto bastante importante se alcança como resultado. Que aspecto é esse? Assinale a alternativa correta.
 - a. Média aritmética.
 - b. Variância amostral.
 - c. Amplitude de valores.

- d. Tabelas de valores.
- e. Padrões.
3. Nos momentos iniciais de produção de grandes massas de dados ou *big data*, utilizou-se um termo para um conjunto de ferramentas que garantiam a capacidade de integração de grandes massas de dados, de análise e de apresentação do tratamento executado. De qual termo estamos nos referindo? Assinale a alternativa correta.
- a. *Big data*.
- b. *Business intelligence*.
- c. *Business analytics*.
- d. Visualização de dados.
- e. Análise de dados.

► Referências Bibliográficas

ACCENTURE. **Entendendo a visualização de dados.** 2014. Disponível em: https://www.accenture.com/_acnmedia/PDF-45/Accenture-Entendedo-De-Dados.pdf. Acesso em: 07 jul. 2019.

ESTADÃO. **Não só de grandes empresas vive o big data e analytics.** São Paulo, 20 maio 2018. Disponível em: <https://economia.estadao.com.br/blogs/sua-oportunidade/nao-so-de-grandes-empresas-vive-o-big-data-e-analytics/>. Acesso em: 08 jul. 2019.

FREITAS, C.M.D.S. et. al. Introdução à visualização de informações. **RITA**, v. III, n. 2, Porto Alegre, 2001. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/19398/000300210.pdf>. Acesso em: 07 jul. 2019.

GRANT, Robert. **Data visualization:** charts, maps and interactive graphics. Boca Raton: CRC Press, 2019.

UNDERS SERVERS E DATACENTERS. **Confira as 8 melhores ferramentas de visualização de dados!** 2018. Disponível em: <https://blog.under.com.br/confira-as-8-melhores-ferramentas-de-visualizacao-de-dados/>. Acesso em: 07 jul. 2019. Não paginado.

Gabarito

Questão 1 – Resposta: C.

Resolução: Grant (2019), em sua publicação, afirmou que William Cleveland escreve que as visualizações traduzem números em imagens.

Feedback de reforço: Lembre-se dos autores citados na leitura fundamental.

Questão 2 – Resposta: E.

Resolução: A análise gráfica de grandes volumes de dados, quando realizada com as ferramentas apropriadas, ajuda a identificar padrões e tendências nos dados.

Feedback de reforço: Lembre-se dos tipos de variáveis existentes nos estudos de análise de dados.

Questão 3 – Resposta: B.

Resolução: Quando as empresas passaram a se preocupar com a produção das grandes massas de dados, passaram a buscar ferramentas que juntas garantiam a integração de dados, a análise e apresentação de resultados. Tais ferramentas são denominadas de *Business Intelligence*.

Feedback de reforço: Lembre-se do significado prático de *business intelligence*.



A organização visual (Visualização de dados e *Big Data analytics*)

Autor: Marcelo Tavares de Lima

► Objetivos

- Apresentar os principais conceitos de *Big Data*.
- Apresentar os principais conceitos de *analytics* e *Big Data*.
- Apresentar os principais tipos de visualização de dados.

1. Introdução

O gerenciamento de negócios baseado unicamente em planilhas é primitivo e desvantajoso de forma ampla nos atuais momentos da humanidade. Todos os setores do negócio podem ser prejudicados por conta dessa maneira “primitiva” de gerenciamento de dados. Tal gestão não acompanha as decisões à velocidade em que dados são produzidos e tratados para gerar informações. Essa é apenas uma das possíveis consequências!

Big Data, no sentido literal da palavra, significa grande volume de dados. No entanto, o sentido prático e real do termo não se limita a isso, é muito mais! Além de se referir a grandes volumes de dados, também, se refere à variedade de dados manipulados das diversas fontes disponíveis e à velocidade em que são tratados.

Além das características citadas, Taurion (2013) acrescenta ainda mais duas variáveis importantes, quando se trata do conceito de *Big Data*, devem ser consideradas a veracidade dos dados (têm significância ou são apenas sujeira?) e o valor para o negócio. E, por último, um elemento que tem estado em voga, a questão da privacidade dos dados, tema complexo e controverso, segundo o autor. Uma regulamentação sobre proteção de dados pessoais foi criada em 2018, a lei nº 13.709/2018 – a Lei geral de proteção de dados pessoais – que tem gerado investimentos nas empresas para proteger e garantir proteção a dados pessoais.

Neste texto serão apresentados conceitos e definições fundamentais associados a *Big Data* e *analytics*, associados com a questão da visualização de dados. Desejamos que você, ao final desta leitura, se familiarize com o tema para que possa se interessar em se aprofundar no tema.

2. Conceitos fundamentais de *Big Data*

Não há um consenso quando se trata do conceito básico de *Big Data*. Por exemplo, Taurion (2013, n. p.) apresenta a definição apresentada pela McKinsey Global Institute como:

a intensa utilização de redes sociais *online*, de dispositivos móveis para conexão à internet, transações e conteúdos digitais e também o crescente uso de computação em nuvem tem gerado quantidades incalculáveis de dados. O termo *Big Data* refere-se a este conjunto de dados cujo crescimento é exponencial e cuja dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados.

Taurion (2013, n.p.) ainda apresenta a definição dada pelo Gartner, declarando que define como *Big Data*.

o termo adotado pelo mercado para descrever problemas no gerenciamento e processamento de informações extremas as quais excedem a capacidade das tecnologias de informações tradicionais ao longo de uma ou várias dimensões.

Uma analogia feita por Taurion (2013) com respeito a *Big Data* e medicina é feita quando o autor afirma que *Big Data* é um microscópio, o qual permitiu que se vissem coisas que já existiam, como bactérias e vírus, mas que não se tinha conhecimento.

A grande massa de dados invisíveis até então, agora pode ser tratada como *Big Data* e receber o tratamento analítico apropriado para se tornarem informações úteis na tomada de decisão. Em era de *Big Data*, com o “microscópio” descoberto, percebeu-se que os dados surgem de todos os cantos. Surgem dos milhares de *web sites* existentes, dos mais de cem mil tuítes por minuto, dos compartilhamentos de bilhões de usuários do Facebook, dos sensores e câmeras espalhados pelas cidades e, não podia deixar de serem citados, dos bilhões de celulares existentes e em funcionamento.

No entanto, o espaço ocupado por esses “novos” dados, que passaram a ser vistos e enxergados como fonte de informação, é maior e, com a produção em massa, requer espaços amplos para o seu armazenamento. Podem ser citados, como exemplo, o uso de imagens e vídeos divulgados nas redes sociais. Um vídeo em alta definição ocupa muito mais espaço para armazenamento em comparação a uma página de texto.

Amaral (2016, n.p.) também apresenta uma definição para *Big Data*, o autor afirma que “*Big Data* é o fenômeno em que dados são produzidos em vários formatos e armazenados por uma grande quantidade de dispositivos e equipamentos”. As causas do “fenômeno” *Big Data*, basicamente, são associadas aos insumos feitos em tecnologia, como, por exemplo, investimentos em unidade central de processamento (CPU), memórias e unidades de armazenamento, tornando-os cada vez mais baratos.

O barateamento dos insumos, a sua miniaturização e o aumento de suas capacidades de processamento têm como consequência a disseminação de mais equipamentos, dispositivos e processos com maiores capacidades de produção e armazenamento de dados. A computação em nuvem e na internet também fazem parte desse pacote de consequências, o que para Amaral (2016) trata-se do que é conhecido como *Big Data*.

A velocidade com que o mundo tem se tornado digital é assombrosa. Na primeira década dos anos 2000, apenas metade dos dados produzidos no mundo estavam armazenados em formato digital e, no final da segunda década (2019), quase 100% dos dados. Então, em notação matemática, Taurion (2013) afirma que *Big Data* = volume + variedade + velocidade + veracidade, tudo agregando valor. É um conjunto de cinco V's. De todos os componentes desta equação matemática, apenas o ‘agregar valor’ não foi discutido.

Aregar valor à implantação de *Big Data* nada mais é do que o retorno esperado de todo o investimento realizado em tecnologia e mão de obra para equipar o negócio. Um conceito que não pode deixar de ser citado quando se fala de *Big Data* é o de internet das coisas (IoT). Taurion (2013, n.p.) afirma que “a Internet das Coisas vai aglutinar o mundo digital com o mundo físico, permitindo que os objetos façam parte dos sistemas de informação”.

Com os objetos componentes da estrutura física que nos rodeia gerando dados a todo instante, tem-se, mais do que nunca, um impulsionador poderoso para *Big Data*. Como, por exemplo, quase todos os componentes de um avião podem gerar dados sobre o seu funcionamento, permitindo que sejam utilizados para realizar previsões com respeito à manutenção, evitando que a aeronave fique parada desnecessariamente, gerando custos sem ganhos. Outro exemplo é a automação residencial, a qual realiza integração entre os equipamentos de uma residência, tornando-a uma casa inteligente.

Com relação à disseminação de equipamentos e processos citados por Amaral (2016), o entendimento deste fenômeno pode ser realizado com um breve levantamento histórico, lembrando que há poucas décadas tinha-se a produção de dados centralizadas em *mainframes* e em computadores pessoais.

Na era *Big Data*, os dados passam a ser produzidos por fontes variadas, como “redes sociais, comunidades virtuais, blogs, dispositivos médicos, TVs digitais, cartões inteligentes, sensores em carros, trens e aviões, leitores de código de barra, [...], celulares, sistemas informatizados, satélites, entre outros” (AMARAL, 2016, n.p.). Os dados produzidos por essas fontes diversas têm formato diverso, diferentes velocidades em sua produção e são de volumes variados também. Algo nunca pensado ou vivido antes!

Por *analytics*, entende-se como um termo amplo que engloba processos e tecnologias. O *analytics* é o processo de extração e criação de informações a partir de dados brutos por filtragem, processamento, categorização, condensação e contextualização dos dados (BAHGA; MADISSETTI, 2019). *Analytics* é o tratamento aplicado no dado para transformá-lo em informação útil para ser aproveitado nos processos de tomada de decisão e para divulgação de resultados.

Avalie as seguintes situações práticas: Como uma empresa mantinha disponibilizadas informações sobre seus funcionários na década de 1990? Elas tinham currículos e formulários preenchidos em seus processos seletivos; outras informações em sistemas de folhas de pagamento, em geral, com difícil acesso; e dados de desempenho individual coletados de forma esporádica por um superior hierárquico. E no século 21? Especificamente na segunda década dos anos 2000, pode-se buscar informações sobre um desses colaboradores em redes sociais, seu histórico de uso de internet, seus e-mails; imagens e vídeos elaborados por ele, dentre outras fontes. Todos estes fenômenos, em sua maioria, sempre ocorreram, a diferença é que, de certo tempo em diante, passaram a ser registrados eletronicamente.

Um comparativo entre celulares e computadores da década de 1980, feito por Amaral (2016, n.p.), concluiu que “*Big Data* fica ainda mais compreensível quando falamos em números: um smartphone de hoje tem maior capacidade que o melhor computador de 1985”.

Amaral (2016) ainda apresenta alguns números para dar uma ideia da dimensão do volume de dados existentes, tal como, a quantidade de pessoas com aparelho celular em todo o mundo contam mais de seis bilhões e são em torno de dois bilhões de pessoas utilizando rede social; em torno de três milhões de e-mails são enviados por segundo no mundo, quinhentos milhões de tuítes por dia e por aí vai.

É importante, também, depois de todo um esforço para definir *Big Data*, saber e definir para ter muito claro o que não é *Big Data*. Ficar restrito unicamente em processos de geração de grande volume de dados não é *Big Data*, pois, além disso, o seu conceito também inclui “mudança social, cultural, é uma nova fase da revolução industrial” (AMARAL, 2016, n.p.). Como já dito, *Big Data* é um fenômeno e não uma tecnologia.

Por ser considerado um fenômeno, *Big Data* “envolve o uso de diversos tipos de conceitos e tecnologias, como computação nas nuvens, virtualização, internet, estatística, infraestrutura, armazenamento, processamento, governança e gestão de projetos” (AMARAL, 2016, n.p.).

O armazenamento indiscriminado de dados oriundos de diversas fontes é uma característica muito associada ao conceito de *Big Data*. Em tempos passados, armazenamento representava um custo monetário alto a ser dispensado para tal. Por isso, armazenava-se apenas o dado que era visto possuir algum valor imediato.

► 3. A organização visual

A escolha das ferramentas de visualizações para a grande gama de dados produzidos no mundo atualmente depende dos vários fatores relacionados aos dados tratados e, também, do público que se deseja atingir.

“As visualizações podem ser estáticas, dinâmicas ou interativas” (BAHGA; MADISSETTI, 2019, p. 38, tradução nossa). Visualizações estáticas são utilizadas quando se possui os resultados da análise armazenados em servidores de armazenamento de dados para apenas apresentar resultados. Se a análise for apresentada com atualizações frequentes, por conta de uma dinâmica de atualização dos dados, deve-se apresentar visualizações dinâmicas com gráficos atualizados dinamicamente. E se as visualizações tiverem fonte de dados alimentadas por usuários diversos, o ideal é apresentá-las de forma interativa.

A análise de dados tem como pano de fundo um contexto ou uma história a ser passada. E, para contar essa história, o processo de visualização de dados tem importante papel nessa tarefa, principalmente, na era de *Big Data*.

Segundo Amaral (2016, n.p.), “visualizar dados permite resumir informações, comunicar de forma mais efetiva, compreender, explorar, interpretar e analisar”. Para que cada um destes itens possam ser cumpridos eficientemente, é necessário cumprir as regras de boas práticas na produção de visualizações.

O uso de ferramentas de visualização tem a intenção de ajudar transmitir de maneira fácil e rápida alguma história representada pelos dados tratados. Para que o nosso cérebro faça uma leitura do que a imagem está querendo nos dizer, precisa ter contato com a imagem apropriada para cada tipo de dado e de público alvo.

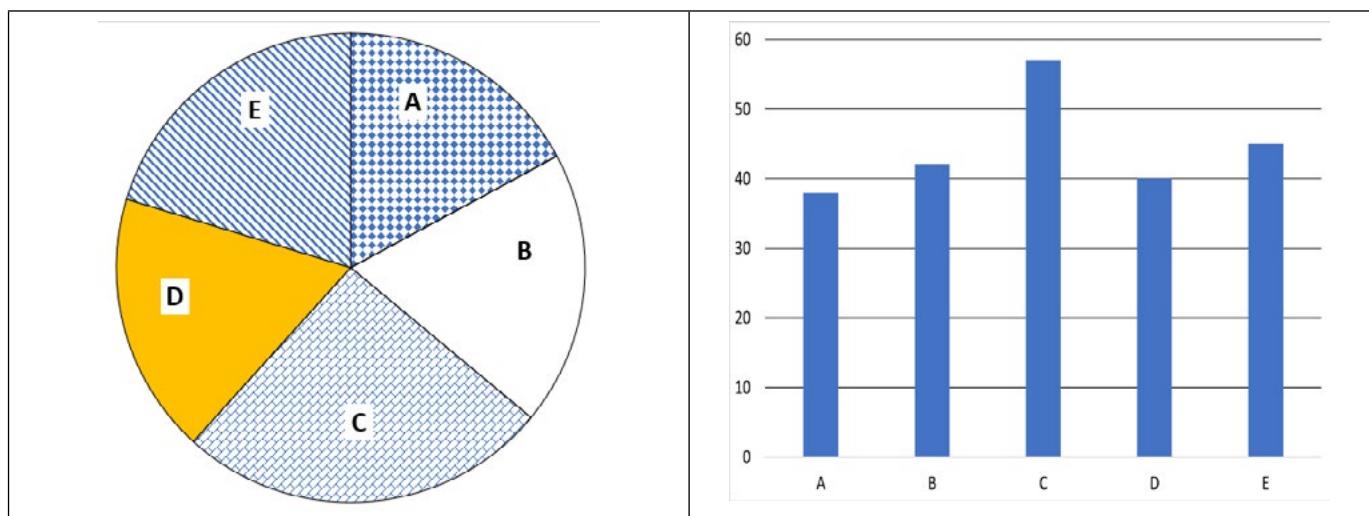
A melhor forma de se começar o estudo sobre visualização de dados é com variáveis quantitativas. Por quantitativa, entende-se o dado que pode assumir qualquer valor numérico dentro de um intervalo de valores. Vale ressaltar que tal conceito está diretamente relacionado com variável quantitativa contínua. A diferença para uma variável quantitativa discreta é que esta tem uma restrição ao assumir valores, pois só pode assumir valores inteiros, como os valores obtidos por contagem ou frequência.

Entretanto, não pode deixar de ser citada, também, o outro tipo de variável comumente encontrada em *Big Data*, a variável categórica, cujo conteúdo representa observação relacionada a uma categoria ou atributo de uma determinada característica.

Amaral (2016) apresenta um teste visual utilizando gráfico de setores (pizza) e gráfico de barras verticais. São gráficos simples e bastante

utilizados. No entanto, quando se pede para que seja feito um ordenamento do maior valor para o menor valor, qual gráfico facilita mais na execução dessa tarefa? Observe a Figura 1 que reproduz os gráficos utilizados no teste.

Figura 1 – Gráficos de setores (pizza) e barras verticais



Fonte: elaborada pelo autor.

Os valores utilizados na elaboração dos gráficos da Figura 1 são diferentes dos valores utilizados pelo autor, no entanto, são proporcionais para que o resultado apresentado aqui seja o mesmo do livro de onde foi retirado.

As imagens foram entregues para voluntários e a eles foi solicitado que ordenassem de forma decrescente segundo as letras de legenda. A ordem dominanteamente atribuída pelos voluntários foi C, B, E, D e A quando viram o gráfico de setores, o qual faz uso de área e ângulo para sua construção. Em seguida, aos mesmos voluntários foi apresentada outra imagem, a qual era um gráfico de barras verticais com os mesmos dados utilizados para a elaboração do gráfico de setores, sendo que para ser construído, o gráfico de barras se utiliza da posição em escala alinhada. O que tinha sido pedido anteriormente, também foi pedido para a segunda figura. Então, a ordem dominanteamente atribuída foi a ordem correta C, E, B, D e A.

Por fim, o que testes como os apresentados nos dizem é que o nosso cérebro ordena melhor dados representados visualmente quando são apresentados em uma escala, preferencialmente alinhada. Em contrapartida, sofre quando se depara com dados apresentados em gráficos que utilizam ângulos ou áreas, pelo menos, quando necessitamos realizar uma ordenação. A lição a ser retirada de testes como este é que temos que entender a importância do uso do elemento visual correto em gráficos.

Ainda assim é possível alguém argumentar que, com um pouco mais de esforço, é possível ordenar corretamente os dados através do gráfico de setores. Mas a intenção é apresentar dados onde o observador tenha facilidade, o máximo possível, em entender a história que está sendo transmitida pelos dados ali representados.

Grant (2019, p. 28, tradução nossa) afirma que “áreas, por exemplo, são geralmente percebidas como mais semelhantes entre si do que realmente são”. O mesmo autor declarou que os pesquisadores em percepção visual humana têm mostrado que não é nada trivial para os leitores de dados apresentados em imagens realizar a conversão para informação numérica.

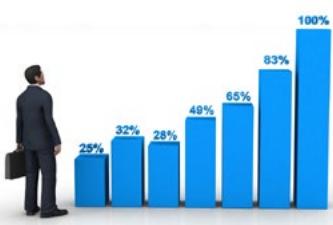
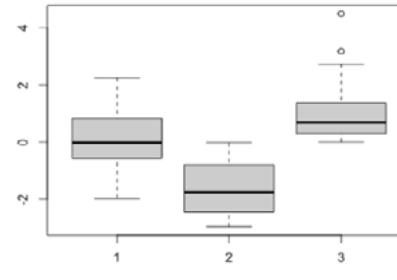
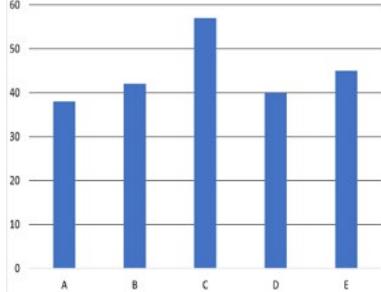
Cleveland e McGill (1984 apud Amaral, 2016) trazem uma relação das dez formas de percepção mais importantes, ordenadas da melhor para a pior, a serem consideradas quando se deseja utilizar recurso visual para apresentação de resultados analíticos de dados. A listagem é replicada a seguir.

- Posição em escala comum.
- Posição em escala não alinhada.
- Comprimento, direção, ângulo.
- Área.
- Volume, curvatura.
- Sobra, saturação, cor.

É claro que o teste com os gráficos de setores e de barras não limita o uso de outros tipos, nem mesmo, torna totalmente inadequado o uso do primeiro tipo de gráfico (setores). É necessário ter boa noção de qual utilizar e como utilizar, fazendo uso de escalas apropriadas, cores devidas, etc.

O Quadro 1 apresenta um guia de qual elemento visual utilizar a depender do tipo de informação que estiver sendo apresentada.

Quadro 1 – Tipos de elementos gráficos adequados

Tipo de gráfico	Quando usar	Exemplo gráfico (*)
Histograma	Mostrar a distribuição de um único dado quantitativo discreto ou contínuo.	 Fonte: Henvry/iStock
Diagrama de caixas (<i>boxplot</i>)	Mostrar a distribuição de um ou mais dados quantitativos contínuos.	 Fonte: Franco, 2012.
Gráfico de barras	Quando envolver uma variável quantitativa (discreta ou contínua) e uma variável categórica (nominal ou ordinal).	 Fonte: elaborado pelo autor.

Séries temporais	Quando dados quantitativos são coletados regularmente em uma escala de tempo.	<p>Fonte: https://en.wikipedia.org/wiki/Paleoclimatology#/media/File:EDC_TempCO2Dust.svg. Acesso em: 05 ago. 2019.</p>
Gráfico de dispersão	Correlacionar duas informações quantitativas.	<p>Fonte: https://commons.wikimedia.org/wiki/File:Gr%C3%A1fico_de_dispers%C3%A3o.png. Acesso em: 05 ago. 2019.</p>
Gráfico de setores	Comparar partes de um total.	<p>Fonte: elaborado pelo autor.</p>

Fonte: Adaptado de Amaral (2016).

Grant (2019) utiliza o termo “*visual encoding of data*” (codificação visual de dados, tradução nossa) para se referir aos parâmetros considerados quando se realiza elaboração de uma imagem para representar uma análise de dados.

Outro parâmetro bastante importante para a elaboração adequada de uma imagem com dados resultantes de uma análise está relacionado com as cores utilizadas para sua construção. Segundo Grant (2019), existem diversas regras para especificar as cores ideais, no entanto, o autor utiliza na maioria dos exemplos apresentados o que ele chama de sistema RGB (*red, green, blue*), ou seja, sistema de cores que utiliza as cores vermelha, verde e azul e suas variações e combinações.

Entretanto, se as cores forem importantes componentes para representação de informações, provavelmente seja importante uma codificação de próprio gosto, respeitando, é claro, a história que a imagem está intencionada em transmitir.

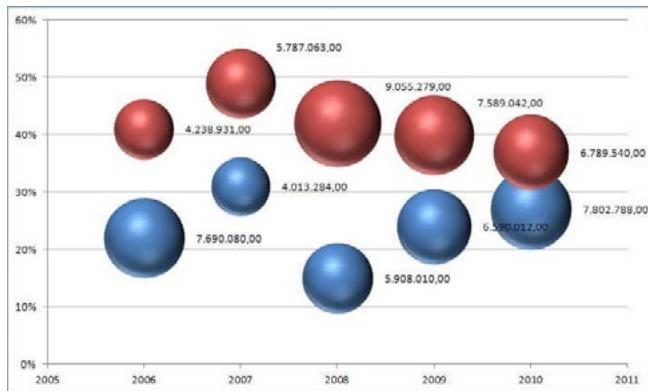
Outro item importante a ser considerado na elaboração de uma imagem diz respeito à representação em áreas, geralmente relacionadas com variáveis quantitativas contínuas. Um exemplo deste tipo gráfico é conhecido como “Bubble chart” ou gráfico de bolhas. É um tipo de gráfico próprio para representar grandezas com valores de áreas ou volumes.

Assim como o gráfico de bolhas, os pictogramas também são representações visuais que precisam de atenção na hora de serem utilizados para representação de informações, pois apesar de serem apropriados para representarem grandezas numéricas, podem confundir facilmente seus leitores.

A Figura 2 apresenta exemplos de gráfico de bolhas e pictograma representando grandezas dimensionais (quantias monetárias e quantidade de lixo recolhido em um período).

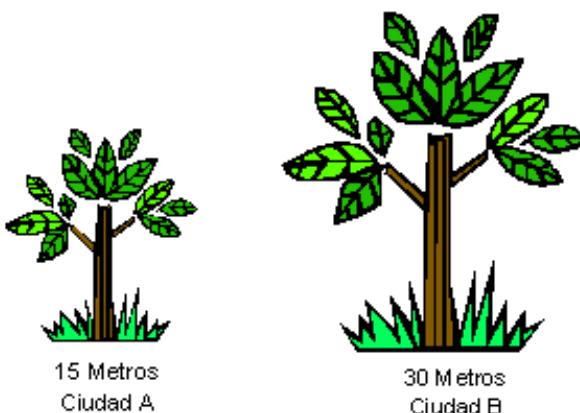
Figura 2 – Exemplo de gráfico de bolhas e pictograma

Gráfico de bolhas



Pictograma

BOTELLAS RECOGIDAS EN UN FIN DE SEMANA



Fonte: Guia do Excel. Disponível em: <https://www.guiadoexcel.com.br/grafico-de-bolhas-excel/>. Acesso em: 10 jul. 2019

Fonte: Representações gráficas. Disponível em: <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/LibroElec/Tema1/pictogramas.htm>. Acesso em: 10 jul. 2019.

PARA SABER MAIS



Como ler uma distribuição de dados: histogramas e gráficos de densidade (kernel) mostram a forma dos dados quando vistos em massa. Se forem uma amostra, podem fornecer uma noção de como os dados populacionais se comportam. Estes gráficos também podem fornecer ideias de algumas medidas estatísticas associadas aos dados, como valores médios, de dispersão, assimetria e moda da distribuição.

Por fim, há de se considerar cada parâmetro de maneira ponderada para elaboração de visualização de dados. No entanto, o responsável pelos dados e pelo tratamento analítico aplicado para a conversão de dado bruto para informação útil é quem terá o poder de decisão final. Ele escolherá o tipo de gráfico apropriado, a cor que lhe convém, etc. O importante é que consiga transmitir o que deseja e que seja

compreendido de forma clara e objetiva, pois a intenção do uso de ferramentas de visualização é a apresentação simplificada dos resultados obtidos de uma análise de dados.

Perguntas que nunca devem deixar de ser levantadas quando se deseja utilizar ferramentas de visualização de dados: (1) Qual a razão para a sua elaboração?; (2) Para quem será criada?; (3) O que se pretende informar?; (4) Qual a melhor forma de apresentação? As respostas ajudarão a encontrar as melhores ferramentas para serem elaboradas com os melhores parâmetros.

ASSIMILE



A visualização de dados pode ser considerada um ramo da ciência da computação, mas também tem relação direta com a estatística e com a ciência de dados. É uma ferramenta que pode ser aplicada em diversos ramos da ciência e de negócios, como, por exemplo, saúde pública, energias renováveis, ciências ambientais e detecção de fraude.

Em tempos de equipes multidisciplinares, um grupo interessado em utilizar recursos de visualização de dados pode ser formado por profissionais diversos, no entanto, para exemplificar, pode ser composto por designers, cientistas de dados, e experts em algum assunto que esteja sendo pesquisado, como, por exemplo, um profissional da saúde, dos esportes, da política, engenheiro ou executivo de empresa, entre outros.

Estamos em época de produção massiva de dados que, após algum tratamento analítico, tornam-se informações importantes que subsidiam tomadas de decisão diversas. Portanto, os conceitos associados a

Big Data, cada vez mais, farão parte do dia a dia dos profissionais de ciência de dados. Dentre as ferramentas disponíveis para apresentação de *Big Data*, uma que se destaca é referente à visualização de dados, que é composta por uma série de conceitos e métodos para aplicar o tratamento apropriado para cada tipo de dado transformado em informação visual.

TEORIA EM PRÁTICA



A visualização de dados é um conjunto de métodos que permitem extrair informações importantes para a tomada de decisão em diversos ramos de trabalho. Para se entender a importância dos métodos de visualização de dados, vale a pena realizar levantamento sobre um pouco da história do *Big Data* para reconstruir um pouco da história da revolução industrial e suas fases na história da humanidade.

Amaral (2016) apresenta de maneira resumida a trajetória histórica das revoluções até chegar a era do *Big Data* e denomina essa era como “A nova onda”.

Suponha que na empresa onde você trabalha é solicitado a você que seja preparada uma apresentação sobre o processo histórico da visualização de dados e do conceito de *Big Data*. Como você faria isso? Por onde começaria? Para ajudar na sua pesquisa, leia a dica do parágrafo a seguir.

O *Big Data* surge em tempos de quarta revolução industrial, trazendo mudanças profundas na indústria, tornando processos produtivos mais eficientes, com menores custos, maior produção e períodos de paradas não programadas cada vez menores.



VERIFICAÇÃO DE LEITURA

1. A representação visual ou gráfica de uma informação deve ser escolhida a depender do tipo de dado utilizado. Uma visualização gráfica que pode representar uma medida numérica e pode assumir qualquer valor numérico dentro de uma determinada faixa de valores. Qual o nome do gráfico que pode representar o conjunto de dados com essa característica?

Assinale a alternativa CORRETA.

- a. Gráfico de setores.
- b. Gráfico de barras.
- c. Histograma.
- d. Diagrama de dispersão.
- e. Gráfico de bolhas.

2. O aspecto de *Big Data* que avalia se os dados gerados e tratados têm alguma significância é chamado de:

- a. Velocidade.
- b. Veracidade.
- c. Significância.
- d. Sujeira.
- e. Valor.

3. Como é classificada uma visualização cujo conjunto de dados que a alimenta é atualizado continuamente por usuários diversos?

Assinale a alternativa CORRETA.

- a. Dinâmica.
- b. Estática.
- c. Visual.
- d. Tabular.
- e. Interativa.

► Referências Bibliográficas

AMARAL, Fernando. **Introdução a ciência de dados:** mineração de dados e *Big Data*. Rio de Janeiro: Alta Books, 2016. KINDLE. Não paginado.

BAHGA, Arshdeep; MADISSETTI, Vijay. *Big Data science & analytics: a hands-on approach*. Arshdeep Bahga & Vijay Madisetti, 2019. ISBN: 978-1-949978-00-1.

FRANCO, G. **Quando usar box plot.** 01/09/2012. Disponível em: <https://sosestatistica.com.br/quando-usar-box-plots/>. Acesso em: 05 ago. 2019.

GRANT, Robert. **Data visualization:** charts, maps and interactive graphics. Boca Raton: CRC Press, 2019.

TAURION, Cezar. *Big Data*. Rio de Janeiro: Brasport, 2013. EPUB. Não paginado. Disponível em: <https://bv4.digitalpages.com.br/#/legacy epub/160676>. Acesso em: 09 jul. 2019.

► Gabarito

Questão 1 – Resposta: C.

Resolução: O histograma é uma representação visual apropriada para mostrar resultados de uma variável quantitativa contínua, ou seja, uma informação numérica que pode assumir qualquer valor numérico dentro de uma faixa de valores.

Feedback de reforço: Relembre o tipo de gráfico utilizado para representar variáveis quantitativas em faixas de valores.

Questão 2 – Resposta: B.

Resolução: O aspecto de *Big Data* que avalia se os dados gerados e tratados têm alguma significância é chamado de veracidade.

Feedback de reforço: Relembre sobre os cinco Vs associados ao conceito de *Big Data*.

Questão 3 – Resposta: E.

Resolução: Uma visualização cujo conjunto de dados que a alimenta e que é atualizado continuamente por usuários diversos é classificada como interativa.

Feedback de reforço: Relembre sobre os principais tipos de visualização de dados.



O processo de Design de *Dashboard*

Autor: Marcelo Tavares de Lima

► Objetivos

- Apresentar conceitos fundamentais de *dashboard*.
- Descrever os diferentes tipos de *dashboard*.
- Descrever os itens necessários e importantes para a elaboração de um *dashboard*.

1. Introdução

A comunicação é uma das principais maneiras de interação e integração entre os seres humanos. Quando bem realizada, ela se torna importante ferramenta para divulgação de ideias e de resultados importantes e decisivos para o sucesso dos negócios e de pesquisas. Por isso, devemos estar sempre em busca da comunicação perfeita.

Um *dashbord* é um dos principais meios de visualização de dados utilizado para comunicar o andamento de trabalhos de uma indústria, de metas de um comércio varejista ou atacadista, dentre outros tipos de informações. Seu uso garante aos envolvidos em um trabalho ou processo o mesmo nível de informação.

Este texto apresenta conceitos fundamentais para a elaboração de um bom *dashboard*, como informações a serem consideradas, os tipos existentes e onde são melhor aplicados, além do *design* que deve ser considerado para uma boa visualização de dados em um *dashboard*.

2. Conceitos fundamentais de *dashboard*

Não há um consenso exato na definição formal de *dashboard*. No entanto, sabe-se que é uma ferramenta de visualização de dados, de maneira rápida e dinâmica, utilizada para apresentar resultados importantes para o acompanhamento de alguma atividade corporativa, de negócios ou acadêmica.

Wexler, Shafer e Cotgreane (2017, n.p., tradução nossa) definem *dashboard* como “um painel ou uma visualização de dados utilizado para monitorar condições e/ou facilitar a comunicação”. Kerzner (2017, p. 255, tradução nossa) afirma que “a maioria das pessoas erram em desconsiderar que visualização de dados é uma ciência, ao invés de arte”.

A definição de *dashboard* apresentada é um conceito, de certa forma, bastante ampla e, por isso, apresentamos os exemplos a seguir para que a teoria possa ser compreendida na prática:

- Uma visualização (painel) interativa que permita aos envolvidos em uma determinada atividade de trabalho, explorar os resultados estratificados por regiões onde atuam, por filial de indústria ou empresa, etc.
- Um arquivo com extensão PDF com resultados de medidores importantes sobre uma determinada atividade de trabalho, o qual é enviado para os executivos envolvidos, toda segunda-feira pela manhã.
- Um aplicativo de celular que permite que os gerentes de venda acompanhem de maneira dinâmica os resultados de vendas.

Um dos principais objetivos de um *dashboard* é a divulgação rápida e clara de resultados através de indicadores e métricas diversas, que possam ser compreendidos por todos os envolvidos em um processo, desde estagiários até os executivos. Métricas segundo Malik (2005, p. 13, tradução nossa) “são medidas de avaliação de performance dentro de um contexto temporal, geográfico e de agregação”.

As informações que devem constar em um *dashboard* estão diretamente relacionadas com os objetivos de uma empresa e, portanto, variam de forma diversa. Portanto, para identificar as informações que devem constar, é necessária a construção de perguntas essenciais que possam direcionar para o que, de fato, é informação importante para ser divulgada e acompanhada periodicamente, para uma correta tomada de decisões associada aos negócios.

As perguntas essenciais e certas para serem acompanhadas via *dashboard* são identificadas quando se conhece as necessidades da empresa de maneira clara e eficiente. Com esta informação, é possível, também, definir as métricas e os indicadores que ajudarão a acompanhar as respostas e os objetivos dos negócios.

Em um instante inicial, pode ser que a(s) pergunta(s) essencial(is) não seja(m) identificada(s). No entanto, não se pode deixar de elaborá-las, mesmo que, em um momento posterior, sejam identificadas como inapropriadas ou imperfeitas. A dinâmica dos negócios também faz com que as perguntas essenciais se modifiquem de tempos em tempos.

Estudiosos do assunto classificam *dashboard* em vários tipos, a depender do tipo de informação que o compõe. De maneira geral, eles apresentam a saúde da empresa. Um *dashboard* que contém informações técnicas, como informações sobre a infraestrutura da empresa, auxiliam na análise de desempenho e de disponibilidade de tecnologias diversas associadas aos processos da empresa.

Quando associado à gestão de negócios, um *dashboard* é composto por um conjunto de indicadores de performance geral da empresa em alguma área específica. Uma das muitas vantagens existentes com a utilização de um *dashboard* é a “libertação” dos relatórios abarrotados de tabelas de dados que tornam cansativo e de difícil interpretação os números contidos neles. Além de simplificar os resultados, um *dashboard* apresenta exatamente o dado que interessa.

Qualquer tipo de dado pode ser manipulado para ser apresentado em *dashboard*. O importante é que sejam dados importantes para a gestão e tomada de decisões, como, por exemplo, dados de estoque, produção por período de tempo, total de vendas de um período, etc.

É comum que os *dashboards* sejam exibidos em grandes telas espalhadas pelos setores envolvidos na produção dos dados, assim como dos executivos dos negócios. Muitos especialistas da área denominam o processo de divulgação e transparência de dados via *dashboard* de “gestão à vista”, pois um *dashboard* ideal, deve conter somente uma página de visualização. O conceito de gestão à vista está

diretamente relacionado com os tipos de métricas e indicadores que serão apresentados no *dashboard*, assim como com os padrões ou mídias visuais escolhidas para a sua elaboração.

Nesse momento, onde está se colocando na prática o conceito de gestão à vista, é preciso ter muito cuidado com a seleção de informações para serem utilizadas na construção do *dashboard*. Não se pode selecionar informação demais, nem de menos. É preciso ponderar adequadamente para não poluir o painel de informações.

Kerzner (2017, p. 255, tradução nossa) afirma que “embora os *dashboards* sejam muito comuns em indústrias, a sua presença pode estar em diversos ambientes”. Por exemplo, um *dashboard* pode ser elaborado para gerenciar o uso de leitos em um hospital. Podem ser instalados em museus, cassinos, consultórios médicos, dentre outros. Enfim, antes de descrever sobre os tipos de *dashboards*: alguns fatos relacionados a eles são elencados, segundo Kerzner (2017):

- *Dashboards* são ferramentas de comunicação.
- *Dashboards* fornecem aos seus usuários o significado atual e futuro de uma informação.
- Quando elaborados apropriadamente, os *dashboards* fornecem informações de *Business Intelligence* (BI).
- Os *dashboards* são relatórios detalhados.
- Alguns *dashboards* podem ser inapropriados para uma situação específica.

A lista de características e/ou aspectos associados a um *dashboard* é longa, ou seja, a lista apresentada anteriormente é muito maior, na realidade. No entanto, ela não será apresentada por completo. A Figura 1 apresenta um exemplo de um *dashboard*.

Figura 1 – Exemplo de um *dashboard*



Fonte: OPSERVICES (2017).

Existem diversos programas que elaboram *dashboards*, que vai desde o MS-Excel, versão 2007 em diante, até R, Python, Qlick Sense, dentre outros. Na internet é possível encontrar muitos manuais de elaboração de *dashboards* e no menu de ajuda dos programas específicos para elaboração dessa ferramenta.

► 3. Tipos de *dashboard*

Não existe um consenso sobre os tipos existentes, principalmente, quanto à nomenclatura criada para a tipologia. No entanto, são aproximadamente semelhantes e, por isso, não há problemas quanto às diferenças existentes.

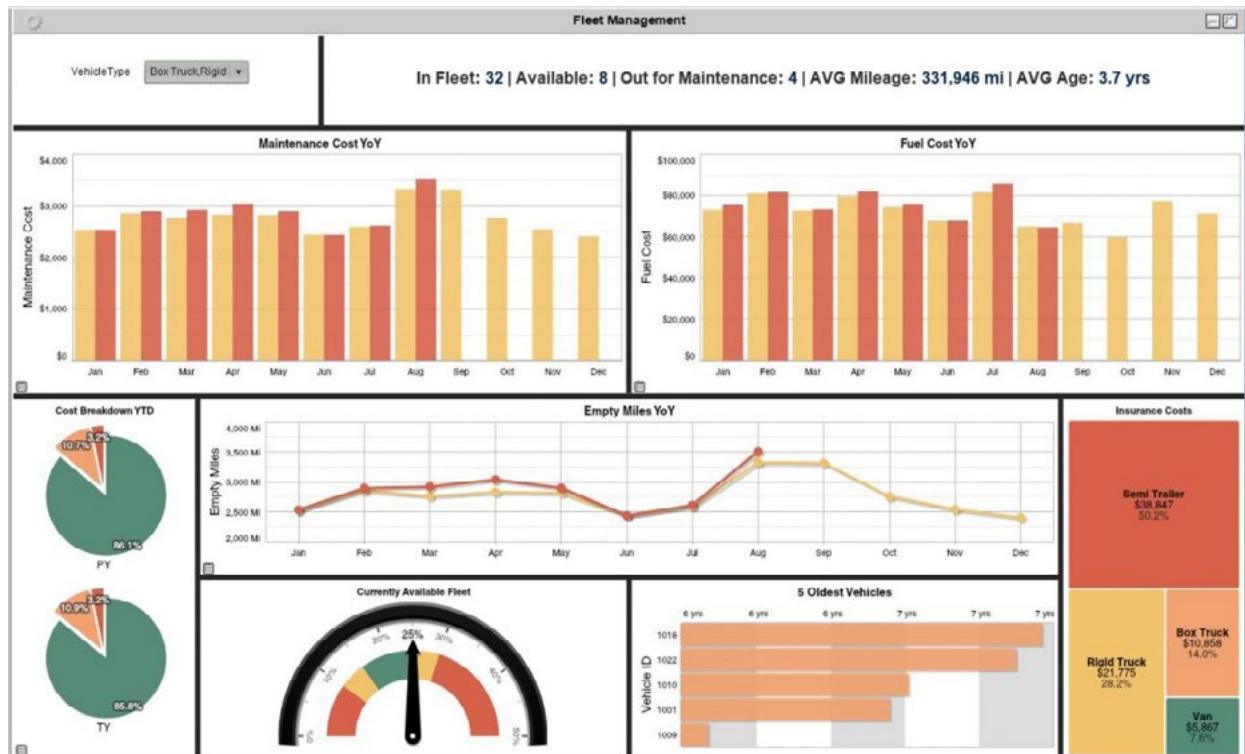
Neste texto serão considerados existentes três tipos de *dashboards*, os quais são denominados como operacional, tático e estratégico, que são os tipos mais gerais. Em termos estéticos pouco diferem entre si. A diferença, de fato, ocorre principalmente quanto ao público-alvo para o qual a ferramenta é elaborada.

3.1 Dashboard operacional

São os que possuem métricas que devem ser acompanhadas para um bom desenvolvimento de uma atividade operacional. São úteis para auxiliar os analistas a corrigir erros e falhas possíveis nos processos de trabalho, os quais poderão ser identificados com maior rapidez através do acompanhamento de um *dashboard*. São painéis com público-alvo, principalmente, os operadores envolvidos nos processos de trabalho. Portanto, é importante aplicar treinamentos adequados em toda a equipe envolvida. Isso acelera o processo de tomada de decisão com as informações disponibilizadas no *dashboard*.

Para exemplificar um *dashboard* operacional, pode-se citar os painéis sobre indicadores de produção de uma indústria, painéis com indicadores de acompanhamento de entrega de mercadorias para acompanhar os atrasos. A Figura 2 apresenta um exemplo de *dashboard* operacional.

Figura 2 – Dashboard operacional



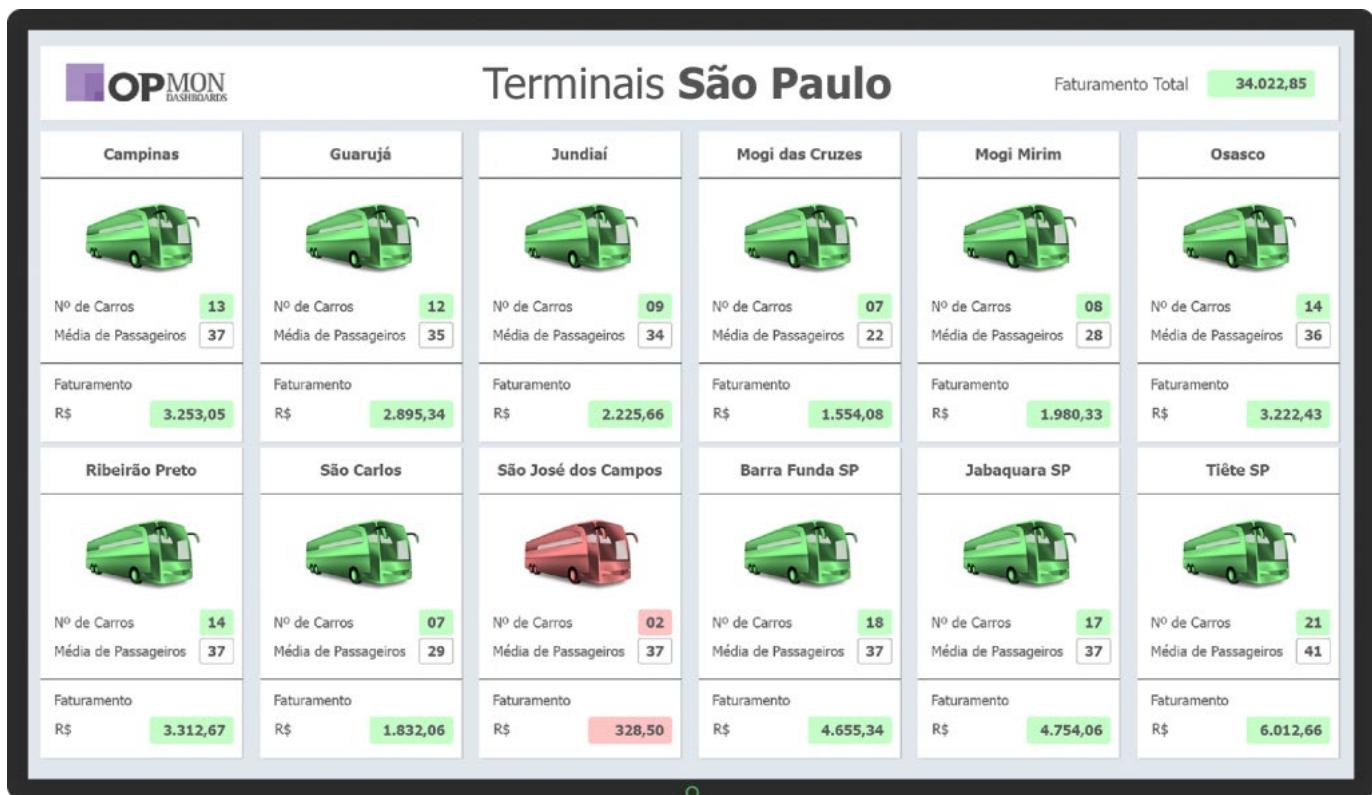
Fonte: <https://www.idashboards.com/blog/2018/08/15/operational-analytical-and-strategic-the-three-types-of-dashboards/>.

3.2 Dashboard tático

São painéis compostos por informações que conseguem permitir que os gestores direcionem recursos para que os objetivos previamente estabelecidos possam ser alcançados em médio prazo. Seu público-alvo principal são as gerências departamentais dos negócios de uma empresa.

Em termos de complexidade, são considerados em maior nível que os *dashboards* operacionais. Por isso, podem influenciar na tomada de decisões para mudanças operacionais de uma empresa. Cada gestor de área pode identificar, por exemplo, gargalos que estejam atrapalhando o processo e, a partir dessa identificação, realizar decisões operacionais. A Figura 3 apresenta um exemplo de *dashboard* tático.

Figura 3 – Exemplo de *dashboard* tático



Fonte: OPSERVICES (2017).

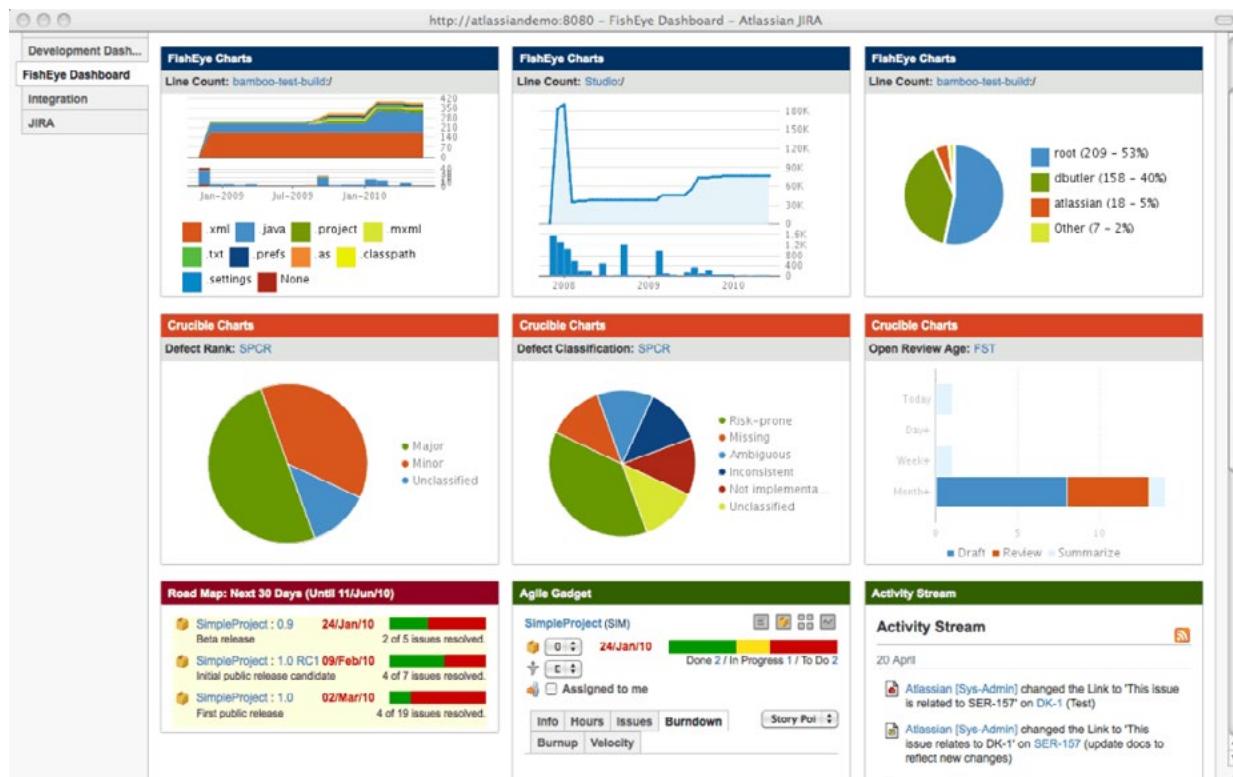
3.3 Dashboard estratégico

Endereçados à alta direção das empresas. Apresentam informações que permitem com que direcionem recursos para que os objetivos, previamente elaborados, sejam alcançados em longo prazo.

Apresentam indicadores que permitem análises comparativas, seja por períodos, regiões ou outras unidades, com a intenção de avaliar a evolução dos trabalhos ou processos, enfim, servem para decisões estratégicas dos negócios.

Apesar de serem direcionados para a alta direção, vale a pena compartilhar, também, com os demais colaboradores da empresa. A intenção é promover maior engajamento nas atividades e aumentar o compromisso com a empresa. Portanto, é uma ferramenta apropriada para ser utilizada na comunicação interna das empresas. A Figura 4 apresenta um exemplo de *dashboard* estratégico.

Figura 4 – Exemplo de *dashboard* estratégico



Fonte: <https://ibid.com.br/blog/kpi/>.

A elaboração de uma ferramenta de visualização de dados como um *dashboard* precisa ser realizada com certo cuidado em relação ao uso de cores, efeitos visuais e quantidade de informações disponibilizadas em seu painel. Esse é um assunto a ser discutido no próximo item.

O Quadro 1 a seguir apresenta um resumo comparativo dos três tipos de *dashboard* apresentados segundo alguns aspectos que os envolve.

Quadro 1 – Resumo comparativos dos três tipos de *dashboard*

Aspecto	Operacional	Tático	Estratégico
Propósito	Monitorar a operação	Mensurar progressos	Estratégia executiva
Usuários	Supervisores e especialistas	Gerentes e analistas	Executivos, gerentes e colaboradores
Escopo	Operacional	Departamental	Empreendimento
Informação	Detalhada	Detalhada/ Resumida	Detalhada/ Resumida
Atualizações	Ao longo de um dia de operação	Diária/ Semanal	Mensal/ Quinzenal
Ênfase	Monitoramento	Análise	Gerenciamento

Fonte: Eckerson, 2017, p. 18.

O Quadro 1 apresenta, de forma resumida, os principais aspectos levantados sobre os tipos de *dashboard* apresentados neste texto em relação aos seus objetivos.

► 4. O processo de *design* de *dashboard*

A Opservices (2017, p. 12) afirma que “a maior parte dos erros na elaboração de *dashboards* diz respeito aos excessos”. O que a empresa quis dizer é que, quando se exagera em alguns itens que compõem a elaboração da ferramenta, isso desvirtua o foco do seu destinatário. Por isso, itens como cores, efeitos visuais e quantidade de informações

inseridas no painel devem ser utilizadas com bastante rigor e moderação.

A visualização de dados exige que transformemos dados em marcas em uma tela. A pergunta que surge é “qual tipo de marca faz mais sentido?” Existe um termo utilizado em inglês, *preattentive attributes* Wexler, Shaffer e Cotgreave (2017), o qual não tem uma tradução exata para o português, mas se refere ao que é primeiramente processado pelo cérebro quando se depara com uma visualização ou painel.

Um *dashboard* bem elaborado é aquele que consegue atrair a atenção de seu leitor exatamente para o que se deseja informar. Além de conseguir ser compreendido, praticamente, de forma instantânea.

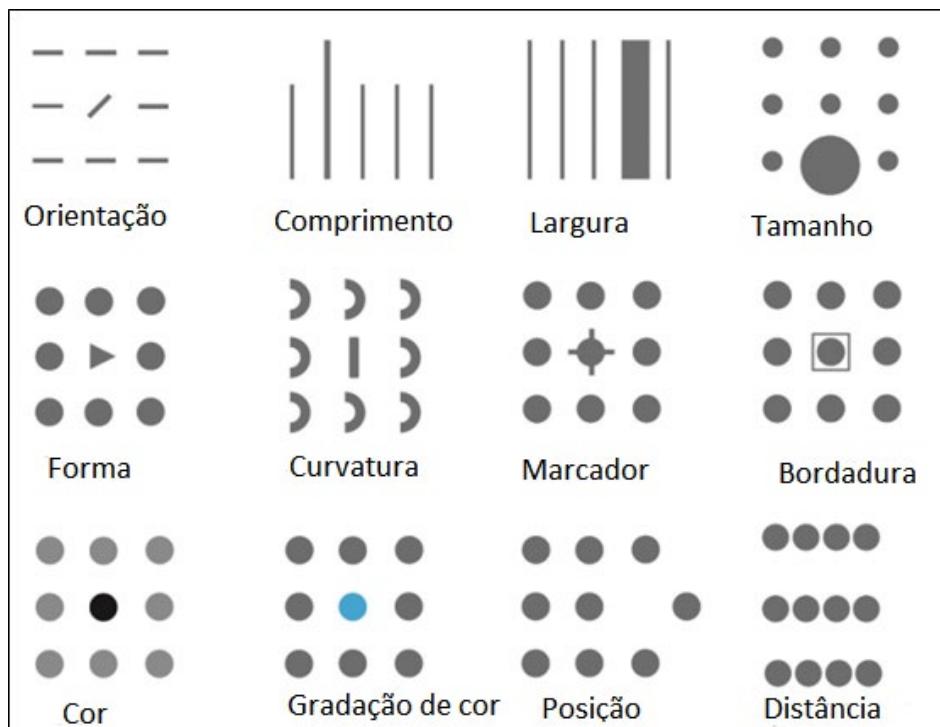
PARA SABER MAIS



É comum achar que os *dashboards* corporativos são apenas para executivos seniores, onde se tem a intenção de lhes dar uma visão geral do desempenho organizacional. Isso não é verdade! A tecnologia de *dashboard* é projetada para ser uma ferramenta eficaz em vários níveis dentro da organização.

Segundo Wexler, Shaffer e Cotgreave (2017), alguns itens que devem ser levados em consideração no momento da elaboração de um *dashboard* e que compõem o que eles denominam de *preattentive attributes* são: (1) orientação das imagens; (2) o comprimento; (3) a largura; (4) o tamanho; (5) a forma; (6) a curvatura; (7) as marcações; (8) as bordaduras; (9) as cores; (10) as graduações de cores; (11) o posicionamento; e (12) a distância entre os elementos. A Figura 2 apresenta um exemplo visual dos doze componentes do *preattentive attributes*.

Figura 5 – Elementos do *preattentive attribute* de uma visualização de dados para elaborar um *dashboard*



Fonte: adaptada de Silverstein (2018).

4.1 Definição das cores de um *dashboard*

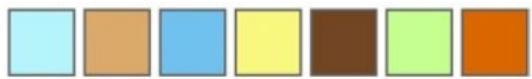
As cores são uma das coisas mais importantes para entender uma visualização de dados e, constantemente, são mal utilizadas. Não se deve utilizar as cores apenas para apimentar uma visualização tediosa. De fato, muitas visualizações excelentes não fazem uso de cores e, mesmo assim, conseguem ser informativas e de aparência elegante.

As cores devem ser usadas de maneira proposital. Por exemplo, podem ser utilizadas para chamar a atenção do leitor quanto a alguma característica, podem destacar parte dos dados ou distinguir entre diferentes categorias.

As cores devem ser utilizadas em visualização de dados de três maneiras principais, segundo Wexler, Shaffer e Cotgreave (2017): sequencial,

divergente e categórica. Além disso, muitas vezes há a necessidade de destacar dados ou alertar o leitor sobre alguma coisa importante. O Quadro 2 apresenta um exemplo desses esquemas de cores.

Quadro 2 – Uso de cores em visualização de dados

Maneira	Visualização
Sequencial: cores ordenadas de mais clara para mais escura. Muito usada em mapas.	
Divergente: duas cores sequenciais com um meio termo neutro. Muito usada em mapas quando se deseja destacar alguma região dos mesmos.	
Categórica: cores contrastantes para comparações individuais. Muito usada em gráficos para destacar as categorias.	
Destaque: cor utilizada para destacar algo. Usada para destacar algum componente de uma imagem.	
Alerta: cor utilizada para chamar a atenção do leitor. Usada para alertar sobre alguma situação identificada na imagem.	

Fonte: adaptado de Wexler, Shaffer e Cotgreave (2017).

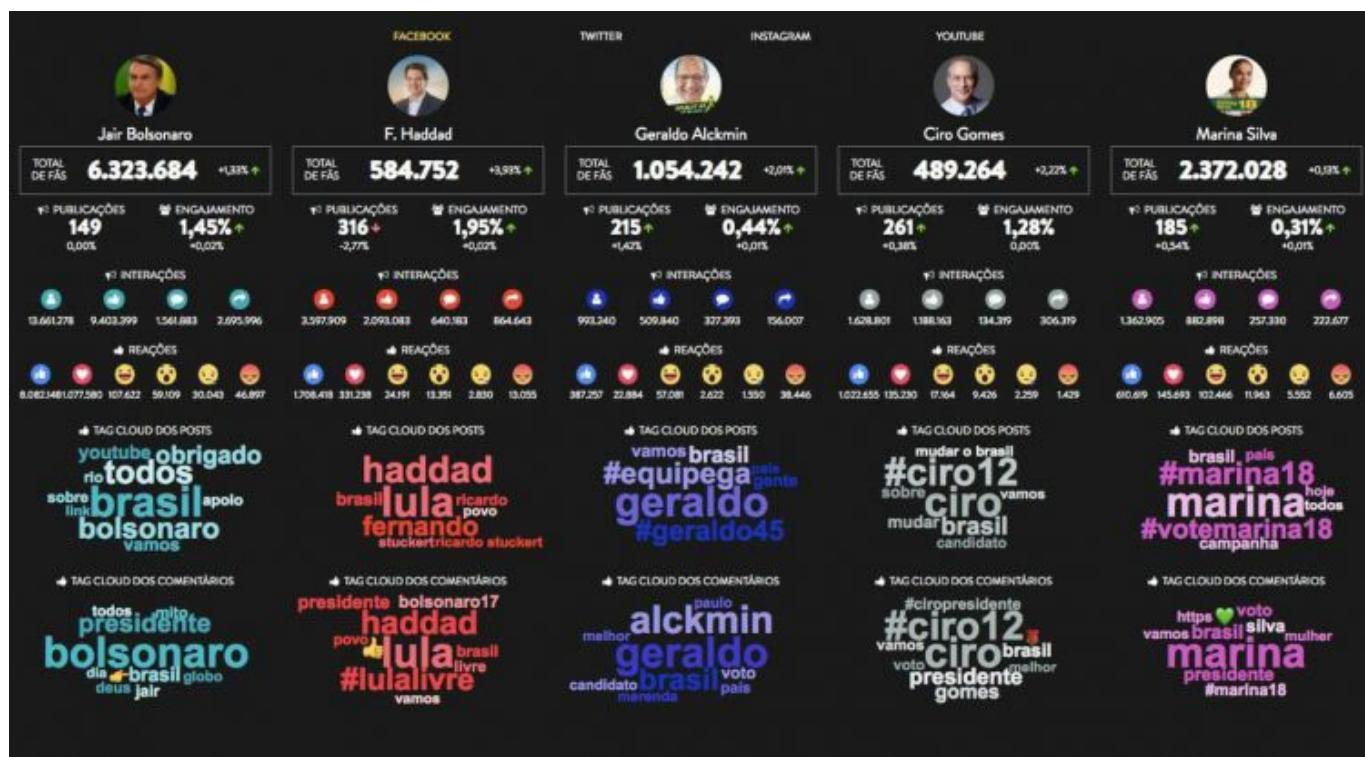
O esquema de apresentação de cores sugeridos por Wexler, Shaffer e Cotgreave (2017) pode ser utilizado em qualquer elemento que for compor um *dashboard*, seja gráfico, mapa ou outro tipo de imagem. O importante é não esquecer que o uso do recurso de cores é o de fornecer um visual agradável aos elementos que compõem o *dashboard*.

4.2 Definição dos efeitos visuais de um *dashboard*

Os elementos que compõem os efeitos visuais de um *dashboard* vão desde gráficos simples e usuais até mapas e outras figuras que possam atrair a atenção do leitor e seja essencial para esclarecer a importância da informação que se deseja transmitir.

Dada a importância deste aspecto, torna-se necessário ser cuidadoso para não pecar por excesso de efeitos visuais, como o uso excessivo de transparências ou de 3D nos gráficos e mapas, pois isso pode tirar o foco do que for realmente importante no *dashboard*. Por exemplo, a Figura 6 mostra um exemplo de *dashboard* com dados sobre a campanha presidencial de 2018, com um certo excesso de informação.

Figura 6 – Excesso de informação em um dashboard



Fonte: <https://blog.zeeng.com.br/category/dashboard/>.

Os efeitos de destaque e de alerta apresentados no Quadro 2 também são itens que compõem os efeitos visuais de apresentação de dados em *dashboard*.

É importante estar atento para a escolha do tipo de gráfico para apresentação de indicadores e resultados. É claro que existem tipos mais apropriados para um tipo de informação ou outra. No entanto, além de estar atento ao tipo de dado a ser utilizado, é preciso, também, lembrar quanto ao contexto no qual o *dashboard* está sendo elaborado e a mensagem que se deseja transmitir com ele, além do público-alvo que se deseja atingir.

O uso de tabelas em um *dashboard* também pode acontecer com algum efeito visual no intuito de atrair a atenção do leitor para algum campo específico do seu conteúdo. Assim como nos gráficos utilizados se deve ficar atento para não cometer exageros na apresentação de tabelas em *dashboards*.

O uso de animações, botões e outros acessórios, também são bastante comuns na elaboração de um *dashboard*. Com estes, os mesmos cuidados devem ser tomados quanto à sua apresentação visual.

ASSIMILE



Lembre-se que as informações que são apresentadas em um *dashboard* são sempre endereçadas a um determinado público, ou seja, um público-alvo. A depender disto é que os *dashboards* foram classificados como operacionais, táticos e estratégicos.

4.3 Definição das informações de um *dashboard*

Quanto às informações que serão apresentadas em um *dashboard*, é preciso ter sempre em mente que um *dashboard* precisa ser

objetivo, ou seja, apresentar apenas o necessário. A definição das informações que compõem o *dashboard* é outro ponto extremamente importante. Um termo técnico bastante utilizado para se referir ao conjunto de informações que compõe um *dashboard* é KPI (*Key Performance Indicators*) que, em português significa indicador-chave de desempenho.

A escolha dos KPIs não difere do processo de escolha de indicadores para elaboração de um relatório ou de um sistema de *Business Intelligence* (BI). Para realizar esta etapa da elaboração de um *dashboard* com rigor e com detalhes, é necessário que o analista envolvido neste processo possua experiência apropriada na área de desenvolvimento.

Esse profissional deve adquirir profundo conhecimento das diferentes fontes de informação dentro da organização e das infraestruturas existentes de *business intelligence* e obter uma compreensão apropriada dos processos envolvidos nos negócios. Além disso, o analista precisa ser capaz de acessar a equipe de especialistas no assunto de divisões como negócios, tecnologia da informação e análise de dados para adquirir uma visão completa do contexto.

Cada KPI selecionado para compor o *dashboard* deve ser dividido em quatro elementos: (1) fonte(s) de dado(s); (2) granularidade; (3) de cálculo e; (4) de variabilidade. Esses elementos juntos definem o escopo completo e esclarecem as diferentes facetas de determinada ação de um KPI.

Este texto apresentou os conceitos fundamentais de *dashboard*, assim como os principais tipos e os principais aspectos associados com o *design* de sua elaboração, como o cuidado com o uso de cores, de recursos visuais e as informações que se deseja utilizar na sua composição.



TEORIA EM PRÁTICA

Dashboard executivo de vendas. O exemplo aqui apresentado foi retirado de Wexler, Shaffer e Cotgreave (2017). Você é um gerente de vendas e quer saber como você e sua equipe têm se desempenhado durante um trimestre específico. Você deseja ser capaz de ver, a qualquer momento durante o trimestre avaliado, como as vendas se desempenharam em comparação a qualquer trimestre anterior. Você precisa ter uma visão global do negócio e também estratificado pelas linhas de produtos individuais ou unidades regionais. Portanto, deseja elaborar um *dashboard* para se manter informado, assim como a sua equipe. Para saber por onde iniciar, é necessário elaborar algumas perguntas, tais como: Como estamos indo neste trimestre? Como está este trimestre em relação ao último trimestre e no mesmo trimestre do ano passado? Estamos no caminho certo para superar o trimestre anterior? Estamos no caminho certo para superar o mesmo trimestre do ano passado? Quais são as transações mais recentes?

Cenários Relacionados

- Os gerentes de produto gostariam de comparar as vendas acumuladas de diferentes produtos lançados em momentos diferentes.
- Os gerentes da área de marketing podem querer medir o quanto viral suas campanhas eram. Quais campanhas receberam o maior número de acessos mais rapidamente? Quais tinham a maior longevidade?
- Os rastreadores de eventos recorrentes usariam *dashboard* como o que pretende elaborar para ver se as suas vendas estão acima ou abaixo da meta em comparação com eventos anteriores.

Como as pessoas deverão usar o *dashboard*?

Este *dashboard* deverá ser projetado para fornecer uma visão geral completa das vendas para dois produtos. Os executivos da empresa receberão uma cópia dos indicadores por e-mail semanalmente. Se precisarem de mais detalhes, eles poderão clicar em um link para ir para a versão interativa visível em um navegador.



VERIFICAÇÃO DE LEITURA

1. Um *dashboard* é uma das muitas ferramentas de visualização de dados existentes. No entanto, existe uma característica peculiar que ele possui. Qual é esta característica?

Assinale a alternativa CORRETA.

- a. Muito colorido.
- b. Divulgação rápida e clara.
- c. Possui muita informação de dados.
- d. Possui somente tabelas.
- e. Acessível apenas aos gestores.

2. Para elaborar um *dashboard* é necessário pensar no seu conteúdo, ou seja, naquilo que ele trará de informação relevante para o seu público-alvo. Dentre os possíveis conteúdos de um *dashboard* está um bastante importante que apresenta resultados numéricos. Assinale a alternativa que contém o conteúdo citado de um *dashboard*.

- a. Figura.
- b. Quadro.
- c. Métrica.
- d. Textos.
- e. Logotipo.

3. Sabemos que um *dashboard* é uma ferramenta essencial para divulgação de informações de forma rápida e eficiente. Seu objetivo é manter os envolvidos em um processo, atualizados e informados sobre qualquer item relevante para o bom andamento do trabalho. Como é conhecido o processo de divulgação e transparência de dados realizado via *dashboard*?

Assinale a alternativa CORRETA.

- a. Painel.
- b. Métrica.
- c. Divulgação interna.
- d. Gestão à vista.
- e. Registro de informação.

► Referências Bibliográficas

ECKERSON, W.W. **Performance Dashboards**: Measuring, Monitoring and Managing Your Business. Hoboken, NJ: John Wiley & Sons, 2006.

KERZNER, H. **Project management metrics, KPIs, and dashboards**: a guide to measuring and monitoring project performance. 3 ed. New Jersey: Wiley, 2017.

MALIK, S. **Enterprise Dashboards**: Design and Best Practices for IT. New York, NY, EUA: John Wiley & Sons, Inc, 2005.

OPSERVICES. **Boas práticas para construir dashboards:** o guia definitivo do assunto. 2017. Disponível em: <http://materiais.opservices.com.br/obrigado-e-book-dashboards>. Acesso em: 16 jul. 2019.

SILVERSTEIN, L. **The science of data visualization.** Tableau Conference, New Orleans, 2018. Disponível em: https://tc18.tableau.com/sites/default/files/session/assets/18BI-030_ScienceOfDataVisualization.pdf. Acesso em: 16 jul. 2019.

WEXLER, S.; SHAFFER, J.; COTGREAVE, A. **The big book of dashboards:** visualizing Your data using real-world business scenarios. New Jersey: John Wiley & Sons, Inc., 2017.

Gabarito

Questão 1 – Resposta: B.

Resolução: Uma das principais características de um *dashboard* é a divulgação rápida e clara de resultados.

Feedback de reforço: Lembre-se do objetivo de um *dashboard*.

Questão 2 – Resposta: C.

Resolução: Dentre os diversos possíveis conteúdos que um *dashboard* pode conter, as métricas são conteúdos que apresentam resultados numéricos.

Feedback de reforço: Lembre-se do conceito de métrica.

Questão 3 – Resposta: D.

Resolução: O processo de divulgação e transparência de dados via *dashboard* é denominado de gestão à vista.

Feedback de reforço: Lembre-se do conceito de divulgação de informações.



Visualização de dados com R, Python e Qlik Sense

Autor: Marcelo Tavares de Lima

► Objetivos

- Apresentar exemplos de visualização de dados com o R.
- Apresentar exemplos de visualização de dados com o Python.
- Apresentar exemplos de visualização de dados com o Qlik sense.

1. Introdução

Estudar a teoria da visualização de dados é extremamente importante para fundamentar o conhecimento sobre esta, que não é uma arte, mas, segundo Kerzner (2017), é uma ciência.

O estudo da prática e exemplificações sobre o tema ajuda o leitor a compreender as aplicações possíveis da teoria de visualização de dados. Ajuda também, a dar ideias ou *insights* para novas maneiras de resolução de problemas e de elaboração de visualizações diversas.

Este texto apresenta exemplos de visualização de dados com o uso da linguagem R, linguagem Python e da ferramenta Qlik Sense. Desejamos que você possa aproveitar bastante esse momento.

Bons estudos!

2. Visualização de dados com a linguagem R

Segundo Wickham e Golemund (2017, p. 1, tradução nossa), “é uma excelente escolha começar com a programação R, pois a recompensa é clara: é possível construir gráficos elegantes que ajudam a compreender os dados”.

O programa R é composto por vários pacotes, os quais também são compostos por funções, sendo estas compostas por linhas de comando. Um dos pacotes construídos especificamente para trabalhar com visualização de dados é o ggplot2, o qual foi desenvolvido por Wickham e Golemund (2017).

A ideia para o desenvolvimento do pacote `ggplot2`, segundo Oliveira, Guerra e McDonnell (2018), embora tenham ocorrido modificações, é originária de uma obra chamada *The Grammar of Graphics*, a qual trata de maneiras de descrever um gráfico a partir de seus componentes". Com o `ggplot2`, segundo os autores, é possível construir gráficos mais complexos de maneira mais fácil.

O pacote `ggplot2` é estruturado para que a "gramática" dos gráficos possa ser utilizada para a elaboração de um gráfico a partir de várias camadas, as quais podem ser formadas por dados, mapeamentos estéticos, transformações estatísticas dos dados, objetos geométricos (pontos, linhas, barras, etc.) e ajustes de posicionamento. Outros componentes possíveis, como os sistemas de coordenadas (cartesiano, polar, mapa, etc.) e as divisões do gráfico também poderão ser utilizados.

Para utilizar o programa R para elaborar visualização de dados é necessário conhecer um pouco de linguagem de programação. No entanto, para facilitar o seu uso é possível trabalhar com a interface gráfica RStudio, a qual tem versão gratuita, assim como o R.

A linguagem de programação R, conforme definem Oliveira, Guerra e McDonnell (2018, p. 10), pode ser entendida como "um conjunto de pacotes e ferramentas estatísticas, munido de funções que facilitam sua utilização, desde a criação de simples rotinas até análises de dados complexas".

A interface gráfica RStudio ajuda ao iniciante em linguagem de programação R a se familiarizar com a construção dos códigos e a inserção de informações necessárias para executar seus comandos. Apesar de ser um facilitador para o uso de linguagem R, o RStudio

também tem uma série de funcionalidades que melhoram o uso da linguagem R. Portanto, também é utilizado por usuários avançados.

O R é inteiramente gratuito e o RStudio tem versão gratuita e versões pagas. Para baixar ambos, iniciando pelo R, basta realizar uma busca na internet e escolher a versão adequada para o seu computador. Para baixar o RStudio, também basta acessar um buscador na internet, como o Google, e procurar a página do programa e escolher a versão adequada para o seu computador.

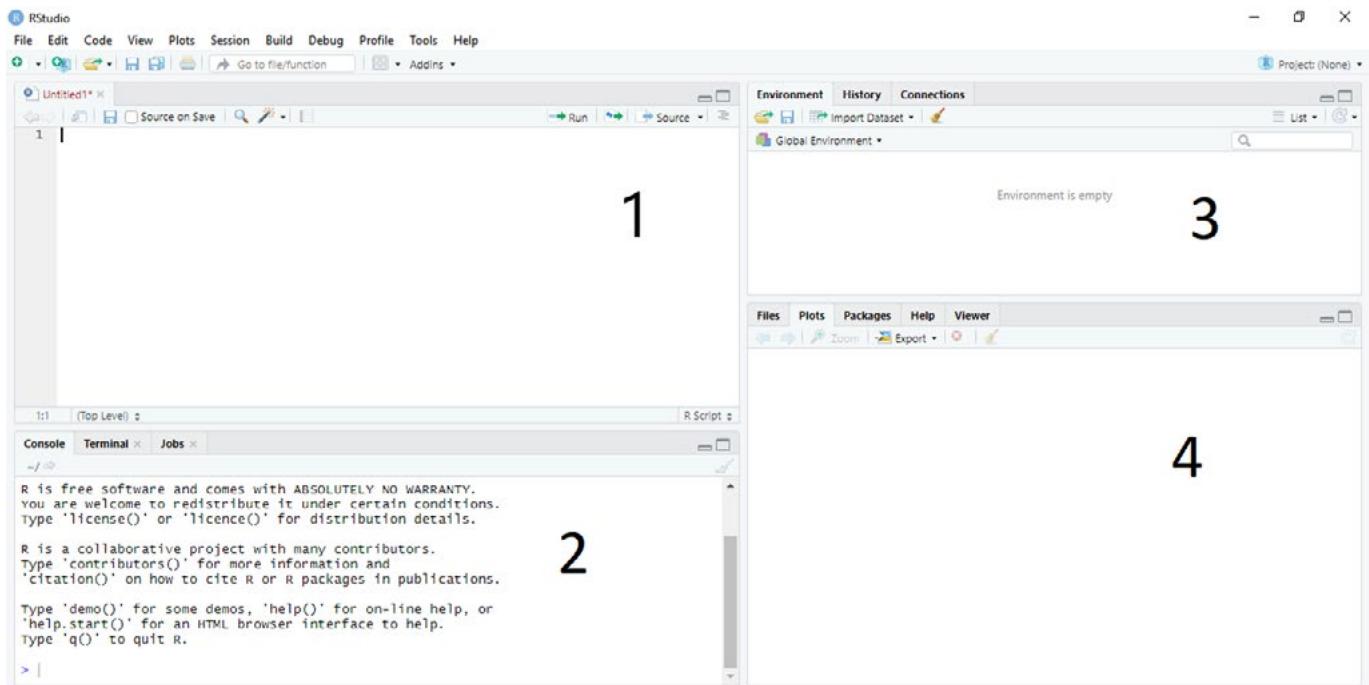
Para realizar a instalação, deve-se seguir as instruções mostradas em sua tela após clicar nos arquivos de instalação baixados em seu computador ou consultar manuais disponíveis na internet. Para utilizar este material com êxito é importante que seja concluída a etapa de instalação dos dois programas.

Tanto o R quanto o RStudio foram, originalmente, desenvolvidos em língua inglesa. No entanto, o R já possui versão em português, mesmo assim, faz-se necessário ter um mínimo de conhecimento de inglês técnico para sua utilização. A versão utilizada neste texto é a 1.2.1335 para Windows 64 bits.

O RStudio é uma das interfaces gráficas existentes para o R, no entanto, é a mais utilizada atualmente e, também, a mais amigável para o uso da linguagem R. Além disso, ela tem um conjunto de funcionalidades que facilita o uso de comandos e obtenção de resultados, “é o que os especialistas em computação chamam de IDE (Integrated Development Environment ou Ambiente Integrado de Desenvolvimento)” (MELLO; PETERNELL, 2013, p. 24).

A interface do RStudio é dividida em quatro partes principais, como mostrado na Figura 1 e detalhadas a seguir.

Figura 1 – Interface do RStudio



Fonte: elaborada pelo autor.

Parte 1. Editor de códigos: é onde se escreve e edita os códigos de linguagem R. Nesse mesmo espaço são criados os chamados *scripts*, ou seja, uma sequência de comandos que serão executados sequencialmente pelo R.

Parte 2. Console: é onde o R mostra a maioria dos seus resultados, ou seja, é onde são mostrados os resultados dos comandos executados. No console, também é possível escrever linhas de comando.

Parte 3. Environment e History: na aba Environment ficarão armazenados todos os objetos criados em uma sessão do R. Pode-se entender como sessão, valendo também para o RStudio, “o espaço de tempo entre o momento em que você inicia o R e o momento em que finaliza” (OLIVEIRA, GUERRA e McDONNELL, 2018, p. 11). Entende-se como objetos as variáveis declaradas nos comandos. Na aba History é criado um histórico dos comandos utilizados na sessão.

Parte 4. Abas Files, Plots, Packages, Help e Viewer: Nesta parte, estão diversas funcionalidades do RStudio. Por exemplo, na aba Files você poderá fazer navegação de arquivos do seu computador, pois são mostradas algumas pastas do Explorer da máquina que está utilizando, o que permitirá definir o diretório de trabalho do R. A aba Plots mostra os gráficos gerados pelos comandos executados. A aba Packages mostra os pacotes instalados no R, e aba Help possui a documentação dos pacotes instalados e muitos exemplos que podem ajudar na construção de *scripts*.

Voltando a falar do pacote de visualização de dados, o ggplot2, uma forma geral (*template*) da sua estrutura, para efeitos de compreensão, é dada pela programação apresentada no Quadro 1.

Quadro 1 – Estrutura geral do pacote ggplot2 do R

```
ggplot(data =<DATA>)+  
<GEOM_FUNCTION>(  
mapping =aes(<MAPPINGS>),  
stat =<STAT>,  
position =<POSITION>  
) +  
<COORDINATE_FUNCTION> +  
<FACET_FUNCTION>
```

Fonte: adaptado de Oliveira, Guerra e McDonnell (2018).

Para utilizar o pacote ggplot2 no RStudio é necessário realizar a sua instalação, pois ele não faz parte dos pacotes básicos do R. Para isso, basta digitar na parte 2 do RStudio, no Console, os seguintes comandos:

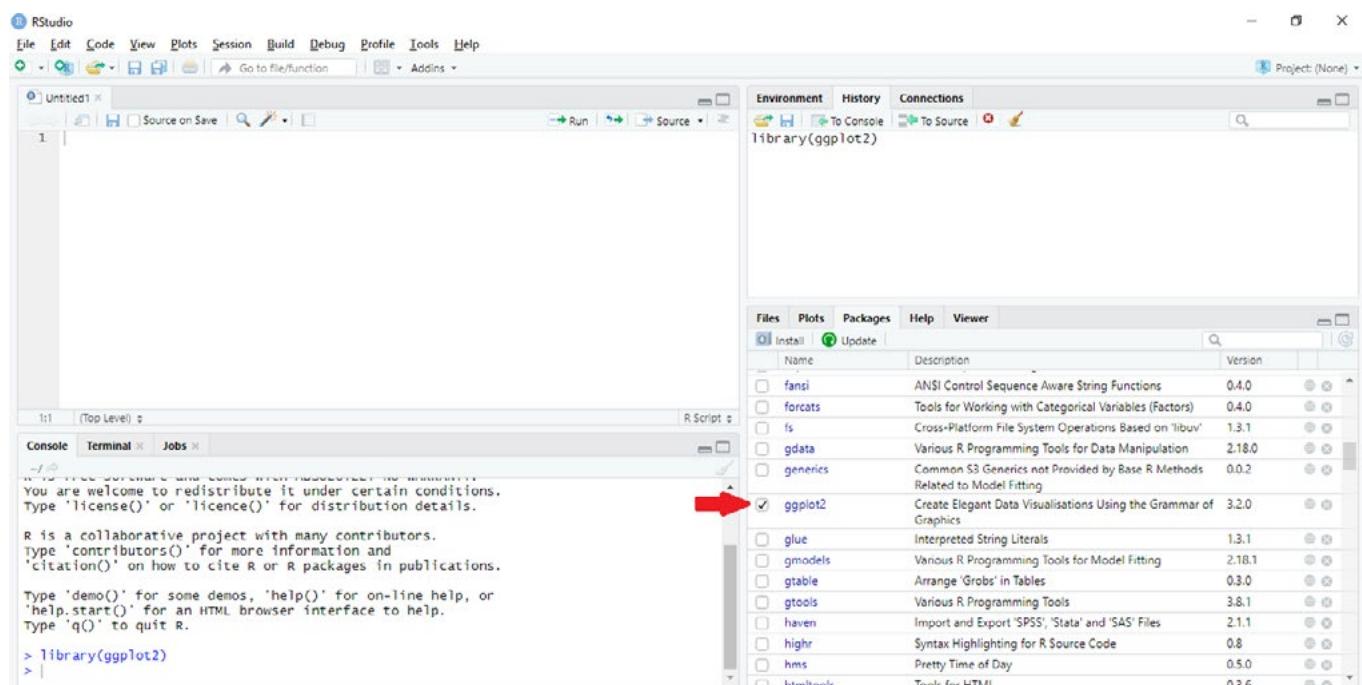
```
Install.packages("ggplot2") # instala o pacote ggplot2
```

```
Library(ggplot2) # carrega o pacote ggplot2 no RStudio
```

O símbolo “#” indica para o RStudio que tudo o que vem em seguida a ele é comentário e não deve ser executado. Não é obrigatório escrever comentários. No entanto, fazê-lo é parte das boas práticas de programação.

Outra maneira muito mais simples de carregar o pacote ggplot2 é através da habilitação do pacote na aba *Packages* da parte 4 do RStudio, como mostrado na Figura 2 (seta vermelha). Basta clicar no botão que fica ao lado do nome ggplot2.

Figura 2 – Carregamento do pacote ggplot2 via RStudio



Fonte: elaborada pelo autor.

Enfim, para elaborar nossa primeira visualização, vamos utilizar um dos bancos de dados disponíveis no pacote ggplot2, chamado mtcars. Não precisa realizar o carregamento do banco de dados, pois ele já é carregado quando faz o carregamento do pacote ggplot2. Os exemplos de visualização que serão apresentados neste texto foram reproduzidos de Oliveira, Guerra e McDonnell (2018).

Para saber o conteúdo do banco de dados basta digitar o comando `?mtcars` no console do RStudio (Parte 2) que aparecerá na parte 4 do RStudio uma descrição sobre o seu conteúdo. No entanto, a descrição na documentação do conjunto de dados afirma que foram extraídos da revista *Motor Trend US* de 1974 e compreendem o consumo de combustível e alguns aspectos de desempenho de 32 automóveis (modelos de 1973 a 1974).

Vamos elaborar um gráfico para nos ajudar a responder às seguintes perguntas: Qual é a relação existente entre consumo e potência do motor dos carros avaliados? Linear? Não linear?

O banco de dados mtcars no R é o que se chama de *data frame*, que é um conjunto retangular de variáveis (colunas) e observações (linhas). Para enxergar o seu conteúdo, basta digitar `mtcars` no console do RStudio.

Para criar um gráfico com os dados do *data frame* mtcars, execute o código a seguir no console do RStudio.

```
# Inicia o plot
```

```
# Adicionar pontos (geom_point) e
```

```
g <- ggplot(data = mtcars) +
```

```
  geom_point(mapping = aes(x = hp, y = mpg))
```

```
# Rótulos(títulos)
```

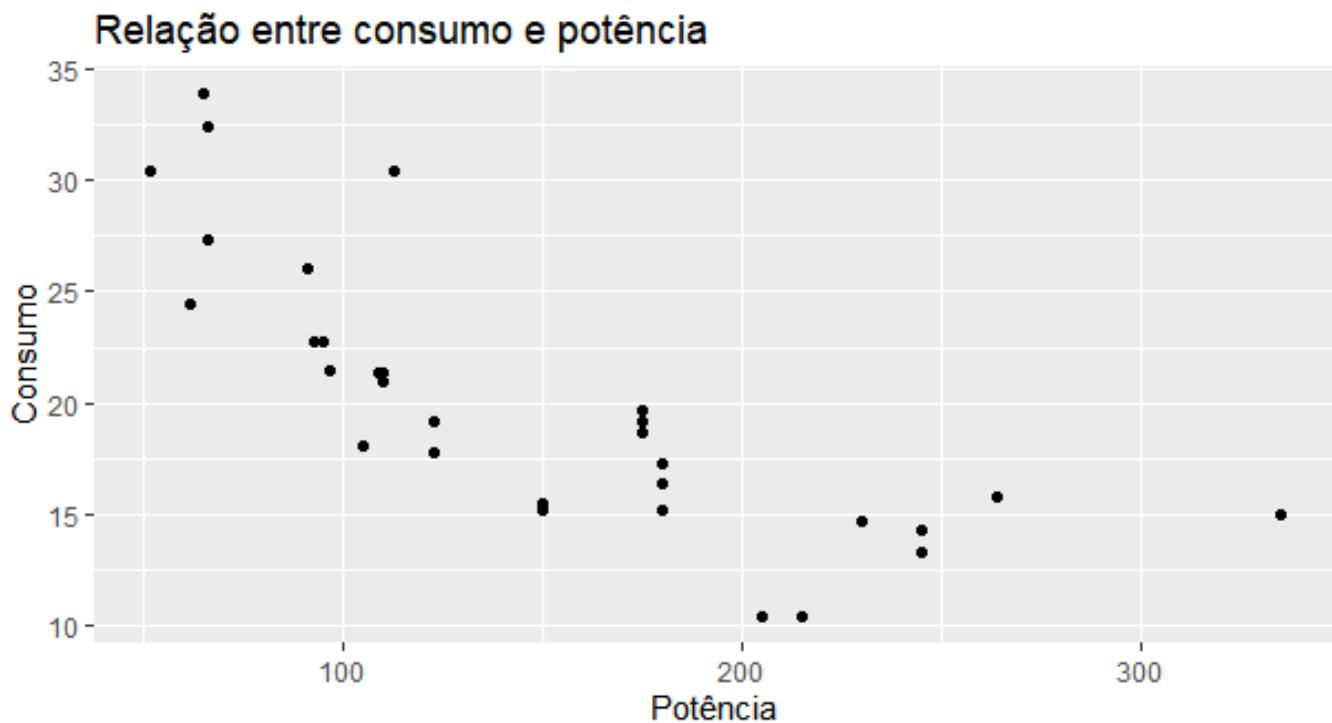
```
g <- g +
```

```
  labs(title = 'Relação entre consumo e potência', y = 'Consumo', x =  
    'Potência')
```

```
g
```

O resultado é apresentado na Figura 3 a seguir.

Figura 3 – Diagrama de dispersão



Fonte: elaborada pelo autor.

É possível observar uma relação negativa entre as variáveis, ou seja, quando uma cresce a outra decresce. Isso quer dizer que quanto maior a potência, menor o consumo de combustível.

Pode-se realizar uma estratificação no gráfico, como, por exemplo, pelo tipo de câmbio (automático ou manual). A programação que deve ser utilizada para isso é apresentada a seguir.

```
g <- g +
```

```
geom_point(aes(x = hp, y = mpg, color = factor(am)), size = 3)
```

```
# Altera a escala dos eixos
```

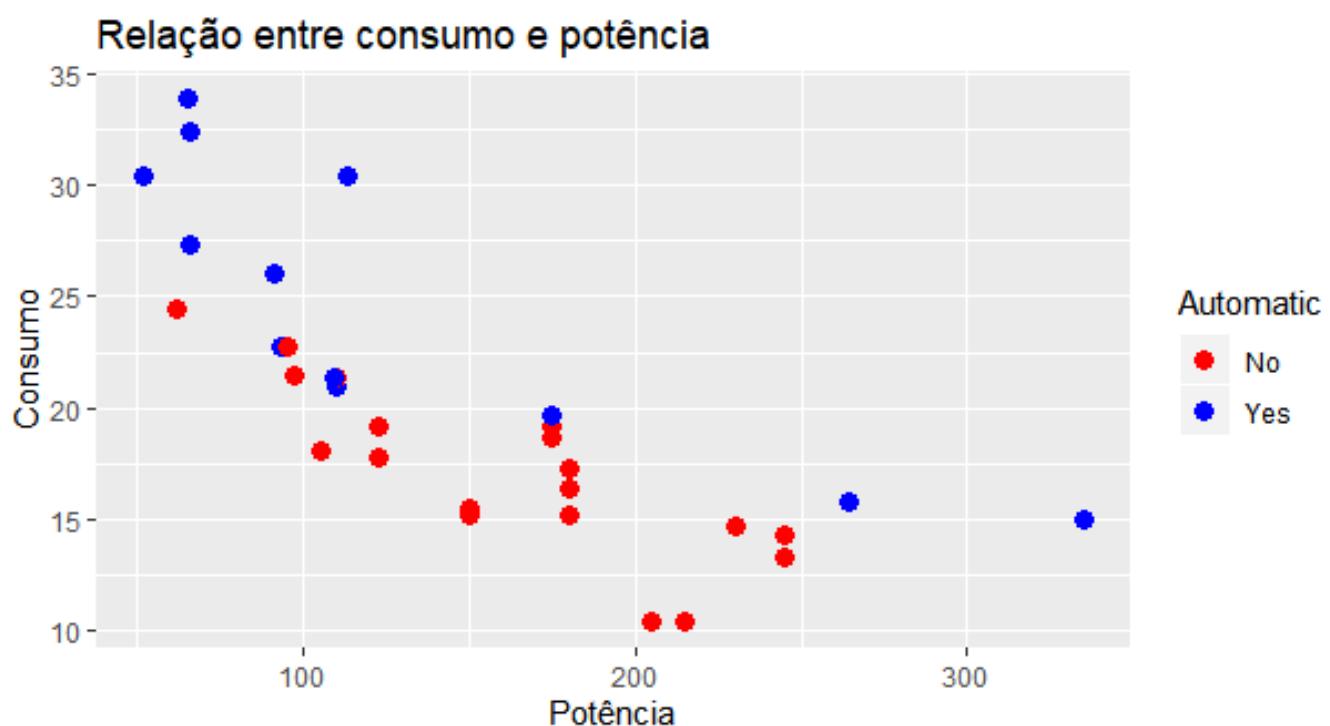
```
g <- g +
```

```
scale_color_manual("Automatic", values = c("red", "blue"), labels =  
c("No", "Yes"))
```

g

O gráfico resultante é apresentado na Figura 4, a seguir.

Figura 4 – Diagrama de dispersão para avaliar a relação entre o consumo de combustível e potência do motor segundo o tipo de câmbio.



Fonte: elaborada pelo autor.

Com esta visualização de dados é possível observar uma série de informações e utilizá-las como subsídio para tomada de decisões. No entanto, vamos nos deter a explicar um pouco do código utilizado.

Primeiramente, foi passado o conjunto de dados mtcars para o ggplot. Em seguida, foi adicionada uma camada de pontos para o mapeamento das variáveis hp e mpg para as posições dos pontos segundo as coordenadas x e y. E, em seguida, realizou-se a estratificação segundo

o tipo de câmbio. Para facilitar a visualização foram adicionados rótulos aos eixos e título ao gráfico.

Existe um mundo de possibilidades para manipulação de dados para produção de visualizações. Existem muitas fontes disponíveis na internet com inúmeras codificações em R para produção de gráficos, desde os mais simples até os mais complexos.

PARA SABER MAIS



R e Excel. É possível importar dados do MS-Excel para serem manipulados no R de várias maneiras. Existem diversos pacotes que fazem a importação, inclusive pacotes que importam dados em rede e grandes bancos de dados também.

► 3. Visualização de dados com a linguagem Python

A linguagem Python foi criada em 1989 pelo pesquisador Guido Van Rossum, do *National Research Institute for Mathematics and Computer Science in Amsterdam* – CWI (SANTOS, 2018). O nome Python foi dado à linguagem por conta de um seriado de comédia que existia na época, cujo nome era Tropa Monty Python.

O Python foi desenvolvido em código aberto (*Open Source*) e a sua primeira versão foi disponibilizada em 1991.

Uma das versatilidades do Python é que ele pode ser programado, tanto no modo funcional estrutural quanto orientado a objetos. Guido Van Rossum tinha como propósito minimizar o máximo possível as

estruturas de programação, retirando as chaves e parênteses excessivos que as linguagens de programação da época continham.

Para baixar o Python é preciso acessar a internet em um buscador, como, por exemplo, o Google, e baixar o arquivo de instalação no seu computador. Basta seguir as instruções que forem aparecendo na sua tela. A versão mais recente para Windows é 3.7.4, para 32 e 64 bits, está datada de 08/07/2019.

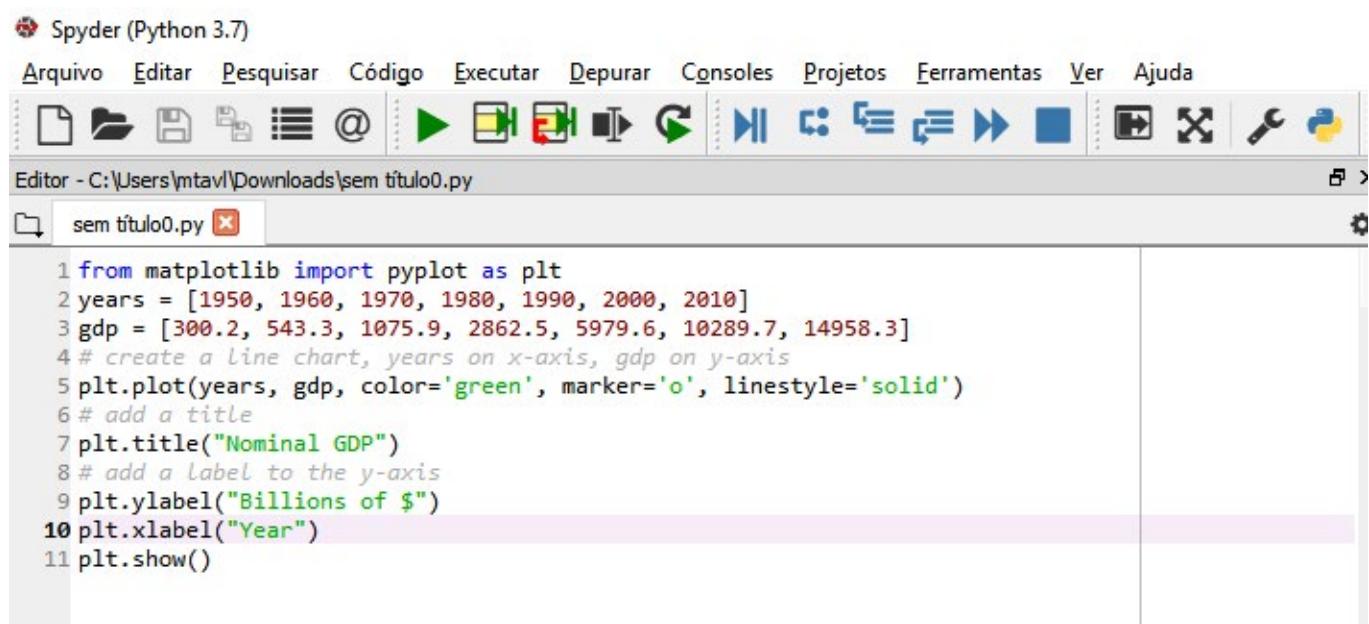
Após o download e instalação do Python 3.7, iremos ter acesso ao IDLE (Python GUI) que é uma interface gráfica básica do programa. Existem interfaces gráficas assim como o RStudio para o R. A interface que será utilizada para o desenvolvimento do programa Python será a Anaconda-Spyder, que pode ser encontrada facilmente na internet.

Em se tratando de gerar visualizações de dados ou gráficos, o Python tem um número razoável de pacotes e bibliotecas para essa finalidade. As principais bibliotecas para gerar e manipular gráficos e imagens são: (1) Seaborn; (2) Matplotlib; (3) Pandas; (4) Altair; (5) Plotly; (6) ggplot e; (7) Bokeh.

Neste texto será utilizada a biblioteca Matplotlib, dentro será utilizado o pacote pyplot, e para começar a utilizar essas ferramentas é necessário realizar as instalações devidas, que podem ser encontradas nos fóruns de internet que tratam do assunto.

Após a instalação do Anaconda-Spyder é possível carregar o pacote pyplot mais facilmente. Para obter um gráfico de linhas, utilizando dados de Grus (2019), por exemplo, a codificação que deve ser utilizada é mostrada a seguir, na Figura 5.

Figura 5 – Linhas de comando de linguagem Python no editor Anaconda-Spyder



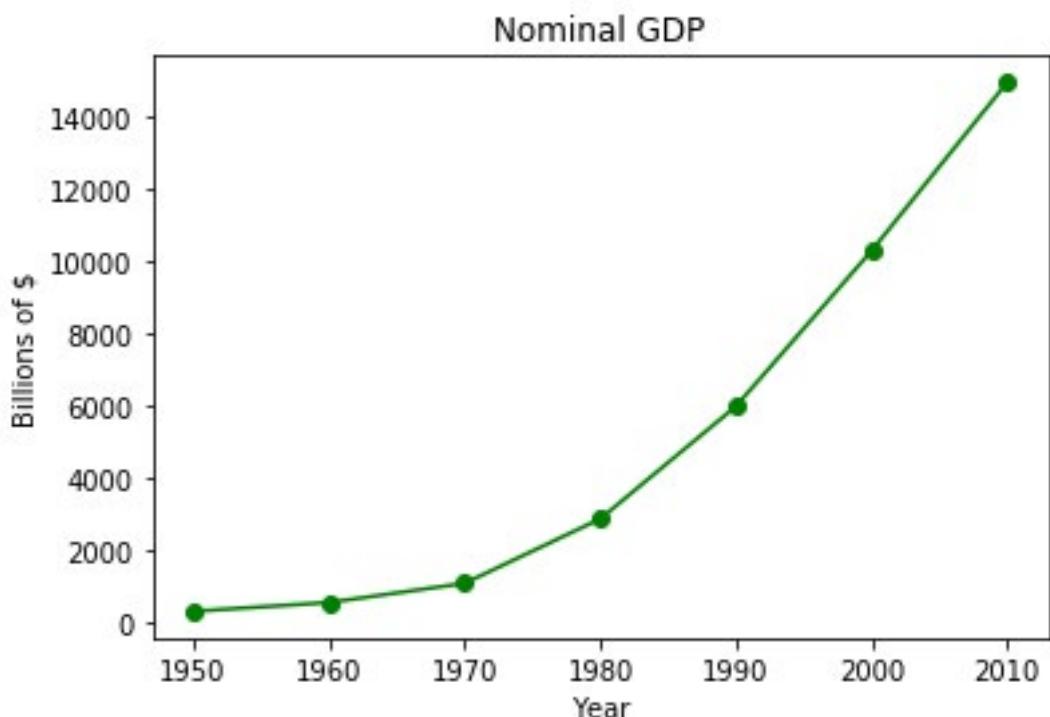
The screenshot shows the Spyder Python 3.7 IDE interface. The menu bar includes Arquivo, Editar, Pesquisar, Código, Executar, Depurar, Consoles, Projetos, Ferramentas, Ver, and Ajuda. The toolbar contains various icons for file operations and execution. The code editor window displays the following Python script:

```
1 from matplotlib import pyplot as plt
2 years = [1950, 1960, 1970, 1980, 1990, 2000, 2010]
3 gdp = [300.2, 543.3, 1075.9, 2862.5, 5979.6, 10289.7, 14958.3]
4 # create a line chart, years on x-axis, gdp on y-axis
5 plt.plot(years, gdp, color='green', marker='o', linestyle='solid')
6 # add a title
7 plt.title("Nominal GDP")
8 # add a label to the y-axis
9 plt.ylabel("Billions of $")
10 plt.xlabel("Year")
11 plt.show()
```

Fonte: elaborada pelo autor.

A visualização resultante é mostrada na Figura 6.

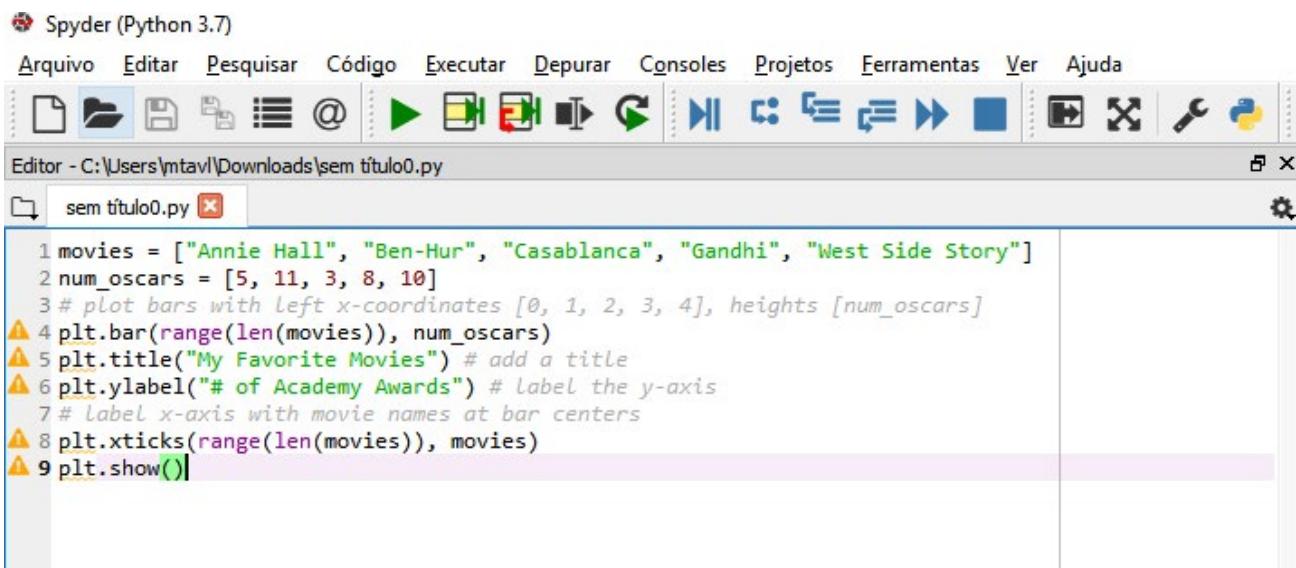
Figura 6 – Gráfico de linhas produzido com linguagem Python



Fonte: elaborada pelo autor.

Para elaborar, por exemplo, um gráfico de barras verticais que envolve a informação de relação entre duas variáveis, onde uma é categórica e a outra é quantitativa, pode-se utilizar as linhas de comando de linguagem Python apresentadas na Figura 7.

Figura 7 – Linhas de comando de linguagem Python para produzir um gráfico de barras verticais



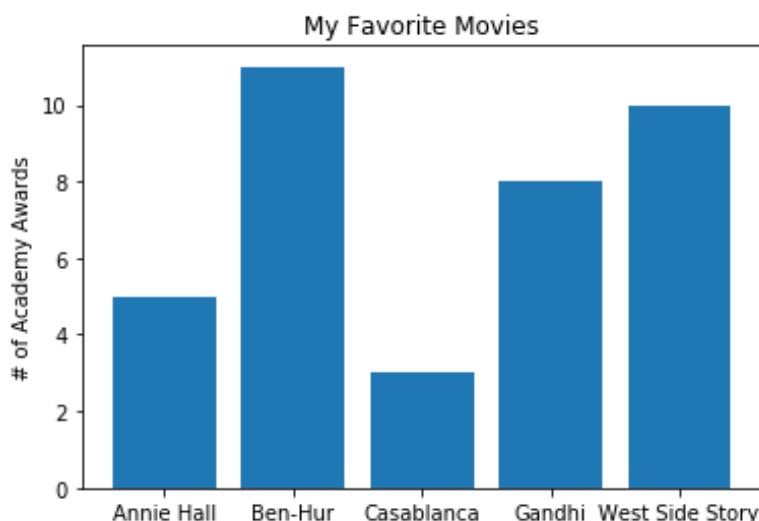
The screenshot shows the Spyder Python IDE interface. The menu bar includes Arquivo, Editar, Pesquisar, Código, Executar, Depurar, Consoles, Projetos, Ferramentas, Ver, and Ajuda. The toolbar contains icons for file operations like Open, Save, and Run. The code editor window displays a Python script named 'sem título0.py' with the following content:

```
1 movies = ["Annie Hall", "Ben-Hur", "Casablanca", "Gandhi", "West Side Story"]
2 num_oscars = [5, 11, 3, 8, 10]
3 # plot bars with left x-coordinates [0, 1, 2, 3, 4], heights [num_oscars]
4 plt.bar(range(len(movies)), num_oscars)
5 plt.title("My Favorite Movies") # add a title
6 plt.ylabel("# of Academy Awards") # label the y-axis
7 # label x-axis with movie names at bar centers
8 plt.xticks(range(len(movies)), movies)
9 plt.show()
```

Fonte: elaborada pelo autor.

A visualização resultante é mostrada na Figura 8.

Figura 8 – Gráfico de barras verticais gerado com linguagem Python



Fonte: elaborada pelo autor.



ASSIMILE

Conhecer linguagem de programação é importante para o profissional que deseja se aprofundar no conhecimento do *business analytics* (BA). É claro que não é obrigatório, mas se torna um diferencial quando ocorre o domínio da mesma.

► 4. Visualização de dados com Qlik Sense

O Qlik Sense é uma plataforma de visualização de dados que permite criar informações interativas para a tomada de decisão. Diferente do R e do Python, o Qlik Sense é um programa pago que não utiliza linhas de comando, criado pela empresa Qlik, a qual iniciou suas atividades em 1993 em Lund, Suécia, mas que permite baixar versões para teste. A sede atual da Qlik fica nos Estados Unidos, na Pensilvânia.

O Qlik Sense Desktop é um aplicativo do Windows que permite a elaboração de relatórios e *dashboards* personalizados e interativos de várias fontes de dados com certo grau de facilidade. Seu uso requer uma conta Qlik.

Para criar uma atividade no Qlik Sense, o que eles chamam de aplicativo, siga as seguintes instruções após instalar e abrir o programa:

1. Acesse o endereço do help online http://helpqlik.com/en-US/sense/June2019/Content/Sense_Helpsites/Tutorials/Tutorials-building-app.htm e clique em “Tutorial – Building an App”. Faça *download* do arquivo compactado. Nele contém um arquivo em PDF com orientações para construir um aplicativo e bancos de dados para serem utilizados como exemplo. Faça a descompactação do arquivo antes de ir para os próximos passos (QLIK TECH INTERNATIONAL AB, 2019).

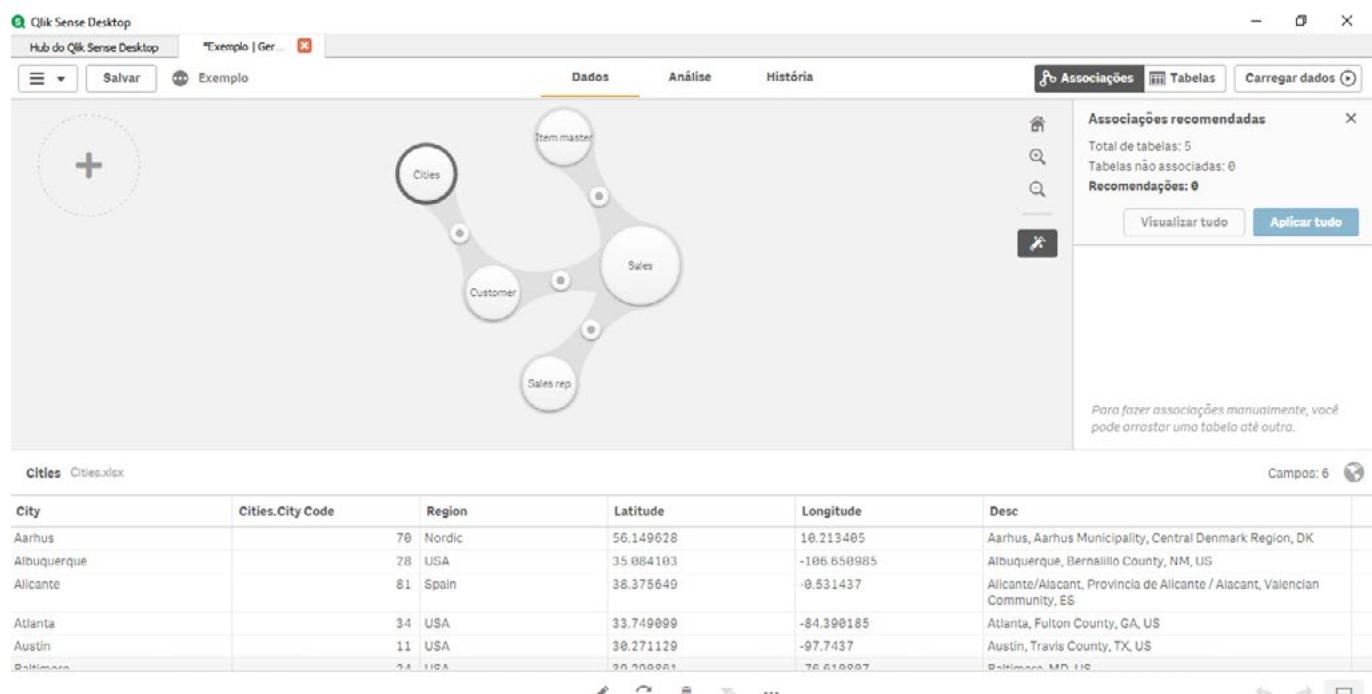
2. No hub (área de trabalho), clique no botão Criar novo aplicativo (canto superior direito da tela). Abrirá uma caixa de diálogo chamada “Criar novo aplicativo” onde há um campo para escrever um nome para o aplicativo denominado “Nome do meu aplicativo:”.
3. Insira o nome do aplicativo.
4. Clique no botão Criar. Um aviso de confirmação de criação do aplicativo aparece na ela.
5. Clique no botão “Abrir aplicativo”. O aplicativo é aberto. Está pronto para receber seus dados.
6. É preciso carregar os dados que serão trabalhados. Iremos utilizar dados disponibilizados pelo próprio programa para treinamento. Iremos carregar os dados da planilha Sales.xlsx. Clique no botão “Adicionar dados de arquivos e outras fontes”. No lado esquerdo da tela, clique em “Meu computador” e busque a pasta onde foram baixados os arquivos.
7. Clique em “Adicionar dados”.
8. Clique no botão “+”, no canto superior esquerdo para associar outros dados.
9. Você deverá buscar na pasta de dados a planilha “Sales rep.csv” e clicar no botão “Adicionar dados”, no canto inferior direito.
10. Arraste a bolha “Sales rep” até a bolha “Sales” para que os dados possam se conectar. O programa detecta as variáveis chave entre os dois bancos de dados, que foi nomeada “Sales Rep Number”.
11. Vamos associar mais dados. Para isso clique no símbolo “+” na parte superior esquerda e selecione da pasta de dados as planilhas “Cities.xlsx”, “Customer.xlsx” e “Item master.xlsx”. Uma de cada vez e, sempre clicando no botão “Adicionar dados”, no canto inferior direito da tela do programa.
12. Para saber como realizar as associações devidas, clique e segure com o botão esquerdo do mouse em cada bolha, como, por exemplo, na bolha “Customer”. O programa mostrará que é possível ligar a bolha “Customer” com a bolha “Sales” ou “Cities”,

deixando verde as bordas dessas duas. O mesmo deve ser feito para as demais bolhas. Para facilitar, é possível clicar no botão “Visualizar tudo”, no lado direito da tela, e depois clicar em “Aplicar tudo”.

13. Clique no botão “Carregar dados”, no canto superior direito e, depois, em “Fechar”.

Após concluir todos esses passos, a tela do Qlik Sense deve ficar igual ao que está mostrado na Figura 9.

Figura 9 – Inclusão de dados no Qlik Sense Desktop

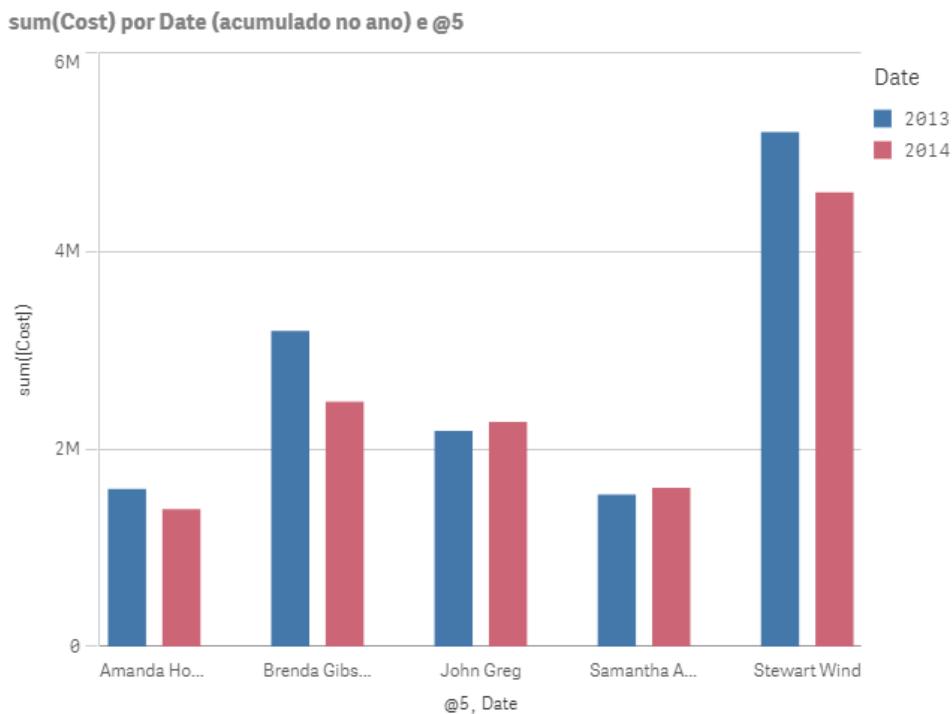


Fonte: elaborada pelo autor.

Uma das vantagens do Qlik Sense Desktop é que ele sugere o tipo de visualização adequada para alguns dados, caso não se tenha uma ideia prévia do que criar.

Por exemplo, clique no botão “Análise” na parte superior da tela e, em seguida, clique em “Gerar ideias”, na parte inferior da tela. Aparecerão uma série de sugestões de gráficos. A Figura 10 é uma das sugestões do programa.

Figura 10 – Gráfico gerado pelo Qlik Sense Desktop



Fonte: Elaborada pelo autor.

É claro que os recursos visuais apresentados neste texto não se esgotam. Existe uma infinidade de possibilidades de produção de visualização de dados, *dashboards*, etc., com os recursos de cada um dos programas aqui tratados. Cabe a você, aluno, buscar o que for de seu maior interesse.

Sugere-se que sejam refeitos os exemplos apresentados e que, leiam, se possível, os documentos de ajuda de cada um dos programas. No início poderá parecer complicado. No entanto, com esforço e dedicação, a curva de aprendizagem logo disparará e você voará no conhecimento das visualizações de dados.

TEORIA EM PRÁTICA



Considere que você deseja elaborar um gráfico de colunas (barras verticais). Esse gráfico parece simples, mas é interessante, pois revela sutilezas. Para exemplificar,

vamos utilizar a linguagem R para elaborar um gráfico de colunas com o pacote ggplot2, extraído de Wickham e Grolemund (2017).

Para isso, será utilizado o banco de dados diamonds, o qual faz parte do pacote ggplot2. Esse banco de dados contém informações de cerca de 54.000 diamantes, como tipo, preço, característica, cor, corte, etc. A programação R utilizada no RStudio para avaliar o tipo de corte do diamante é a seguinte.

```
library(ggplot2)  
ggplot(data=diamonds)+  
  geom_bar(mapping=aes(x=cut))
```

É possível dar uma melhorada no gráfico acrescentando uma programação para melhorar a estética.

```
ggplot(data=diamonds)+  
  geom_bar(mapping=aes(x=cut,fill=cut))+  
  labs(title = 'Distribuição de frequências dos diamantes  
por corte',  
       y = 'Frequência absoluta',  
       x = 'Tipo de corte')
```

Ainda é possível acrescentar uma segunda variável, como, por exemplo, a característica de claridade do diamante, com a seguinte programação.

```
ggplot(data=diamonds)+  
  geom_bar(mapping=aes(x=cut,fill=clarity))+  
  labs(title = 'Distribuição de frequências dos diamantes  
por corte',  
       y = 'Frequência absoluta',  
       x = 'Tipo de corte')
```

É possível manipular muito mais a visualização para que ela fique esteticamente do jeito que se deseja. Para isso, basta incrementar a codificação.

Bons estudos!

VERIFICAÇÃO DE LEITURA



1. A linguagem R é composta por pacotes que são compostos por funções que foram criadas por diversos estudiosos. Cada pacote foi criado para realizar alguma análise, seja produção de medidas estatísticas ou gráficos. Um destes pacotes foi criado para gerar visualizações de dados complexos que podem compor, por exemplo, *dashboards*. Qual o nome do pacote?

Assinale a alternativa CORRETA.

- a. geom_bar.
- b. geom_point.
- c. ggplot.
- d. ggplot2.
- e. mapping.

2. A linguagem Python foi criada por Guido Can Rossum para ser uma linguagem versátil e flexível. Em que ano ela foi criada?

Assinale a alternativa CORRETA.

- a. 1988.

- b. 1989.
- c. 1991.
- d. 1990.
- e. 1992.

3. O Qlik Sense Desktop é uma plataforma do Windows para elaboração de visualização de dados. A construção de uma atividade de visualização de dados no Qlik Sense recebe um nome específico. Qual é este nome?

Assinale a alternativa CORRETA.

- a. Desktop.
- b. Banco de dados.
- c. Gráfico.
- d. Aplicativo.
- e. Programa.

► Referências Bibliográficas

GRUS, J. **Data science from scratch:** first principles with Python. 2. ed. Sebastopol: O'Reilly Media, Inc., 2019.

KERZNER, H. **Project management metrics, KPIs, and dashboards:** a guide to measuring and monitoring project performance. 3. ed. New Jersey: Wiley, 2017.

MELLO, M. P.; PETERNELLI, L. A. **Conhecendo o R:** uma visão mais que estatística. Viçosa, MG: UFV, 2013.

OLIVEIRA, P.F.; GUERRA, S.; McDONNELL, R. **Ciência de dados com R:** introdução. Brasília: IBPAD. 2018. Disponível em: <https://www.ibpad.com.br/o-que-fazemos/publicacoes/introducao-ciencia-de-dados-com-r#download>. Acesso em: 17 jul. 2019.

QLIK TECH INTERNATIONAL AB. **Qlik Sense Desktop, versão 13.32.2.** Qlik, 2019. Online Help.

SANTOS, R.F.V.C. **Python:** guia prático do básico ao avançado. Série cientista de dados, 2018. E-BOOK KINDLE. Não paginado.

WICKHAM, H.; GROLEMUND, G. **R for data Science**: import, tidy, transform, visualize, and model data. Sebastopol: O'Reilly, 2017.

► Gabarito

Questão 1 – Resposta: D.

Resolução: O pacote do R que foi elaborado para gerar visualização de dados de forma elegante é o `ggplot2`.

Feedback de reforço: Lembre-se da palavra *plot*, que significa gráfico em inglês.

Questão 2 – Resposta: B.

Resolução: A linguagem Python foi criada em 1989, mas sua primeira versão foi divulgada em 1991.

Feedback de reforço: Lembre-se que foi no final dos anos 1980.

Questão 3 – Resposta: D.

Resolução: Para criar uma atividade de visualização de dados no Qlik Sense, usualmente se fala que vai criar um aplicativo.

Feedback de reforço: Lembre-se dos programas de celulares.



Visualização de dados utilizando ferramentas OLAP

Autor: Marcelo Tavares de Lima

► Objetivos

- Apresentar conceitos fundamentais de OLAP.
- Apresentar conceitos associados com as ferramentas OLAP.
- Descrever aplicações e visualizações com ferramentas OLAP.

1. Introdução

A importância de se ter boa informação pode ser pensada com o trabalho dispensando e calculado como a diferença de valor entre decisões corretas e decisões erradas, onde as decisões são baseadas nessas informações. Quanto maior a diferença entre as decisões certas e erradas, maior é a importância da boa informação (THOMSEN, 2002).

A tomada de decisões através de visualização de dados garante um bom resultado quando se utiliza de ferramentas técnicas e metodológicas apropriadas para o tratamento devido dos dados de interesse. As ferramentas de visualização de dados multidimensionais, também conhecidas como OLAP (*online analytical process*), são uma dentre as muitas ferramentais disponíveis para a visualização adequada para se obter um bom subsídio na tomada de decisões.

Inicialmente, funcionam como interfaces de visualização de dados, no entanto, vão muito além disso. Podem realizar a manipulação rápida e eficiente de dados complexos e multidimensionais. Desejamos que este texto possa agregar significativamente para você quanto aos conceitos associados a OLAP. Tenha um excelente momento de estudos!

2. Conceitos básicos de OLAP

O conceito OLAP não é um conceito novo, o seu desenvolvimento, segundo Vieira (2009), deu-se início em 1962, através da publicação de um livro denominado *A programming language*, de autoria de Kenneth Iverson, professor de matemática nascido no Canadá.

Os conceitos OLAP mais parecidos com os últimos conhecidos são originários da linguagem APL, desenvolvida no final dos anos 1960 pela IBM. Estes conceitos foram introduzidos nos anos 1990 e, em relação à linguagem APL, tinham maior integração ao acesso de dados.

OLAP (*Online analytical process*), conhecido em português como processo analítico online ou sistema de informações multidimensionais, de maneira prática, é a interface entre a grande massa de dados complexos armazenadas em um banco de dados e o seu usuário.

Outro conceito divulgado sobre OLAP, Inmon (1999 apud Vieira, 2009, p. 16) afirma que “Olap é uma tecnologia de *software* que possibilita uma variedade de visualização das informações que antes era de uma coleção de dados referentes ao empreendimento”. Por conta dessa ampla definição, segundo Thomsen (2002), pode-se falar em conceito OLAP, em linguagem OLAP e produtos OLAP.

Os conceitos OLAP incluem a noção ou ideia de múltiplas dimensões hierárquicas e podem ser usados para se pensar mais claramente sobre a estrutura dos dados.

As linguagens formais OLAP incluem linguagem de definição de dados (DDL), linguagem de manipulação de dados (DML), linguagem de representação de dados (DRL) e analisadores associados (e compiladores opcionais), os quais podem ser usados para qualquer modelagem descritiva, seja transacional ou de suporte a produtos OLAP completos que precisam incluir um compilador e métodos de armazenamento e acesso de dados.

Os produtos OLAP incluem compiladores, métodos de armazenamento, otimização de acesso aos dados e cálculos, utilizados para suporte à decisão e modelagem descritiva de dados (DSS) (THOMSEN, 2002).

Portanto, não se trata de um *software* específico, mas de um conceito onde vários programas computacionais podem ser enquadrados como OLAP, tanto pela capacidade de manipulação de dados quanto pela capacidade de apresentação visual. É uma ferramenta que permite ao usuário usufruir de uma análise multidimensional (MDA),

ou seja, de poder manipular dados que estejam armazenados em dimensões diversas.

Como função básica de uma ferramenta OLAP, pode-se enumerar: (1) visualização multidimensional de dados; e (2) exploração de dados. A etapa do armazenamento de dados está mais vinculada a um outro conceito conhecido como *data warehouse*, que em uma tradução literal significa depósito de dados digitais. Na prática, *data warehouse* é onde são armazenados dados digitais de uma empresa, no intuito de disponibilizá-los para produção de relatórios.

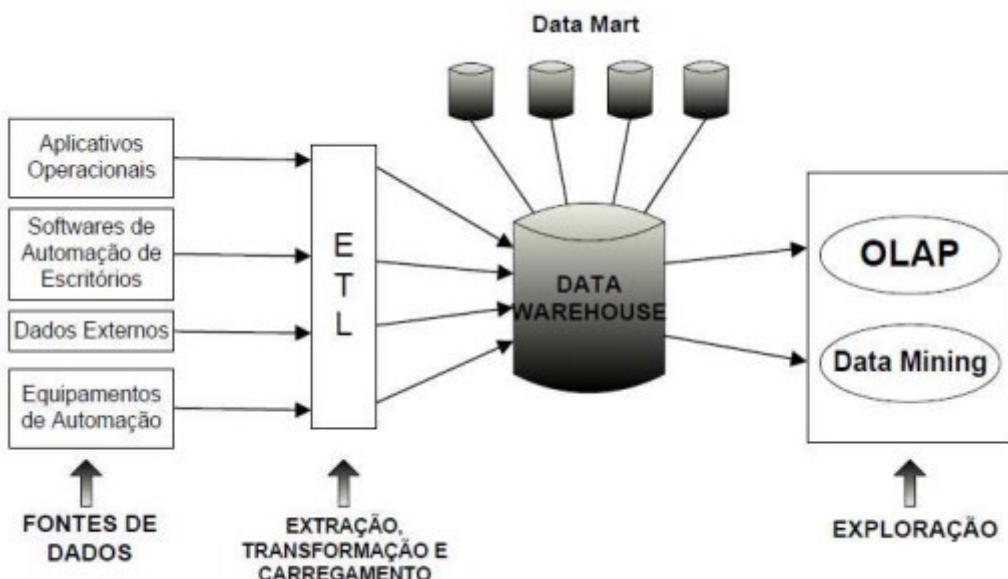
Os conceitos OLAP e *data warehouse* caminham juntos quando abordam o assunto de manipulação e visualização de dados complexos. Caminham juntos porque não se pode falar de um sem abordar o outro.

O conceito de *data warehouse*, de forma simples e genérica, significa depósito de dados orientado por assunto, por período, dentre outros agrupamentos, para dar suporte à tomada de decisões de empresas, negócios, etc., subsidiando o controle de processos e de determinações de padrões. Ainda sobre *data warehouse*, Inmon (1997 apud Vieira, 2009, p. 13) “é um conjunto de dados consolidados por assunto, não é volátil e está sempre em constante variação quanto ao tempo”.

Vieira (2009) apresenta uma listagem de características de um *data warehouse*: (1) conjunto de programas para extração de dados; (2) banco de dados; (3) sistema para recuperação e visualização de dados. Em termos de benefícios, um *data warehouse*, segundo Vieira (2009), podem ser descritos: (1) Acesso fácil e rápido a dados; (2) Armazenamento de dados de forma consistente; (3) Flexível; (4) Detentor de ferramentas de consultas e de visualização de dados; (5) Armazenamento confiável de dados; (6) Construção de dados específicos para direcionamento de negócios.

De forma esquemática, pode-se apresentar um fluxograma dos conceitos de OLAP e *data warehouse* como na Figura 1.

Figura 1 – Fluxograma de OLAP e *data warehouse*



Fonte: Rocco (2009).

Os *data warehouses* permitem acesso amplo e visão multidimensional de vários conjuntos de dados, incluindo cálculos matemáticos e estatísticos mais complexos.

Um outro conceito bastante utilizado quando se trata de OLAP é o conceito de *data mining*, cuja tradução literal é mineração de dados. Em termos conceituais, *data mining* tem por objetivo identificar correlação entre informações com ferramentas estatísticas, matemáticas e computacionais com o intuito de mostrar tendências e padrões.

PARA SABER MAIS



OLAP e *data mining*. Ambos são parte de qualquer processo de tomada de decisão, no entanto, enquanto os sistemas OLAP focam em dados multidimensionais, os sistemas de *data mining* focam, em geral, em dados unidimensionais. Outra diferença entre eles é que nos sistemas OLAP a informação já é conhecida, o usuário

apenas a acessa para produzir algum subsídio para utilizá-la, enquanto que nos sistemas *data mining* o usuário não tem conhecimento da informação ou, tem pouco conhecimento sobre ela.

► 3. Pensando em N dimensões

A noção de hipercubo, um cubo com mais de três dimensões, é fundamental para compreender e trabalhar com softwares de análise multidimensional de dados, que usam dados de planilhas eletrônicas e banco de dados baseado em tabelas. Estes softwares permitem a navegação, geração de relatórios e análises avançadas de dados. Rodrigues et al. (2012, p. 2) afirmam que:

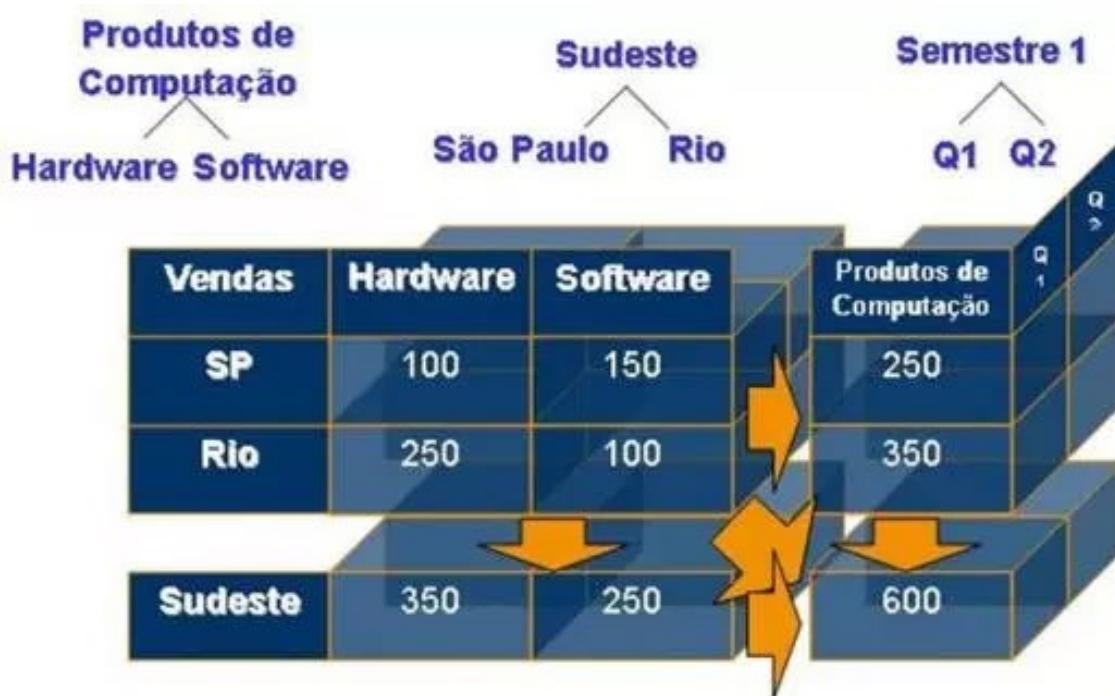
a multidimensionalidade se dá pelo fato de que os dados podem ser visualizados em diversas faces, causando uma ideia de cubo". Cada uma das faces apresenta uma significação, delimitando o assunto que se deseja analisar.

Vamos começar com um exemplo típico de dados bidimensionais, que é a forma mais comum de armazenamento de dados. Qualquer coisa que você rastreie ou faça algum tipo de acompanhamento, seja horas por funcionário, custos por departamento, saldo por cliente ou reclamações por loja, pode ser organizado em um formato bidimensional. Considere que estas informações sejam levantadas mensalmente.

Dados bidimensionais podem ser facilmente armazenados em uma planilha de dados, como o MS Excel, por exemplo. Os meses do ano podem ser inseridos em linhas e as informações diversas podem ser arranjadas em colunas, alinhadas com o respectivo mês do ano.

O que aconteceria se adicionarmos uma terceira dimensão, por exemplo, denominada de produtos? Seria fácil realizar uma visualização? Pensando geometricamente, sim, pois, é apenas um cubo. A Figura 2 mostra um modelo de cubo tridimensional que representa os produtos, as regiões geográficas de atuação e os períodos em semestre. O que esse cubo está nos mostrando?

Figura 2 – Cubo tridimensional de armazenamento de dados



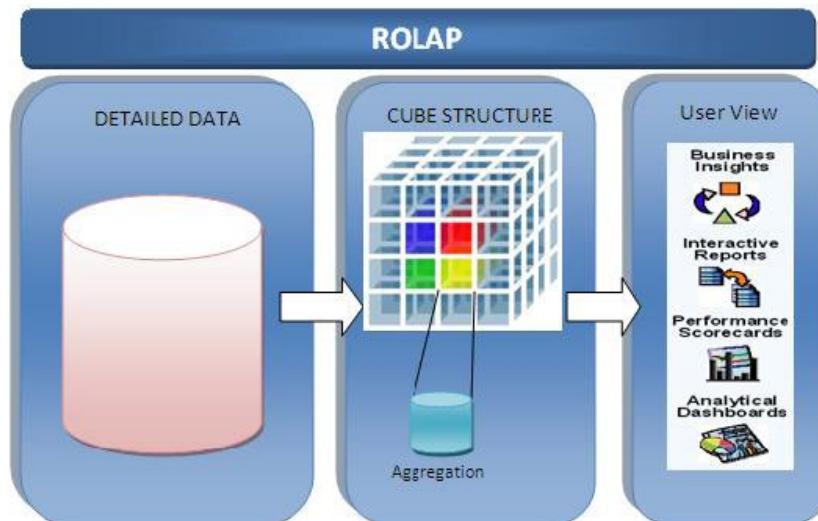
Fonte: Ribeiro (2013).

Obviamente, não é assim que os dados são vistos na tela do computador. Esta é uma representação do armazenamento dos dados, no *data warehouse*. Armazenamentos de dados em cubos ou hipercubos agiliza o processo de busca de informações e de visualizações.

Quando se fala em cubo ou hipercubo, na realidade está sendo referenciado ao que se chama de arquitetura de dados ou *data warehouse*, os quais podem ser classificados como: (1) ROLAP; (2) MOLAP; (3) HOLAP, (4) DOLAP, dentre outros. Uma breve descrição sobre cada tipo de arquitetura ou armazenamento será apresentada a seguir.

- ROLAP (OLAP relacional): a consulta realizada é enviada ao servidor de banco de dados relacional e ali mesmo é processada, mantendo o cubo no servidor de dados. A Figura 3 apresenta um esquema desta estrutura de armazenamento.

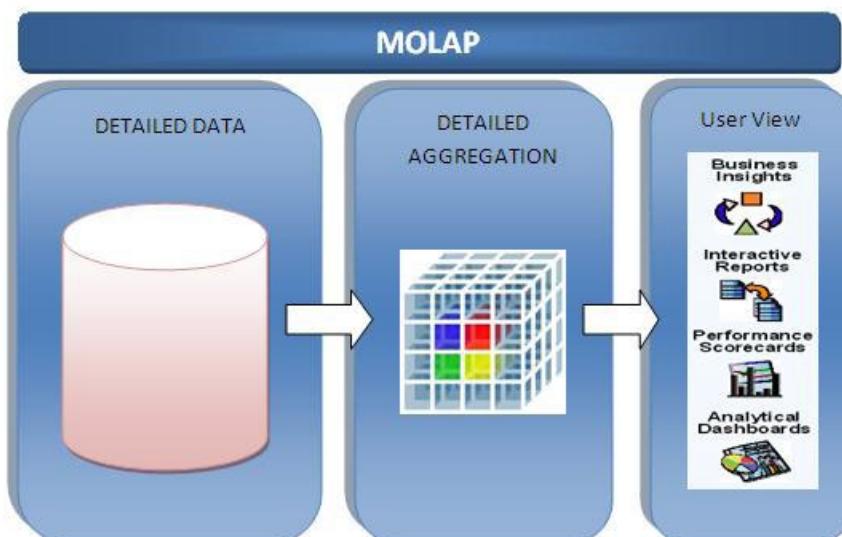
Figura 3 – Arquitetura ROLAP



Fonte: Online Analytical Processing (OLAP) (2009)

- MOLAP (OLAP multidimensional): o armazenamento de dados é realizado de forma multidimensional. Utiliza *data warehouse* com várias dimensões para realizar acesso ao banco de dados. A Figura 4 apresenta um esquema.

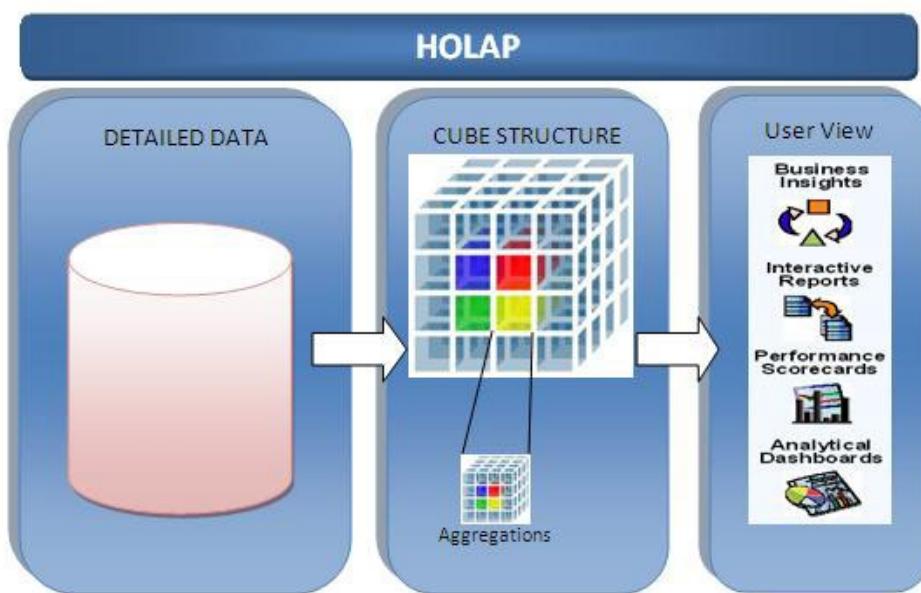
Figura 4 – Arquitetura MOLAP



Fonte: Online Analytical Processing (OLAP) (2009).

- HOLAP (OLAP Híbrido): é uma combinação da arquitetura ROLAP com a MOLAP. É a integração de ambas. A Figura 5 mostra um esquema dessa integração.

Figura 5 – Arquitetura MOLAP



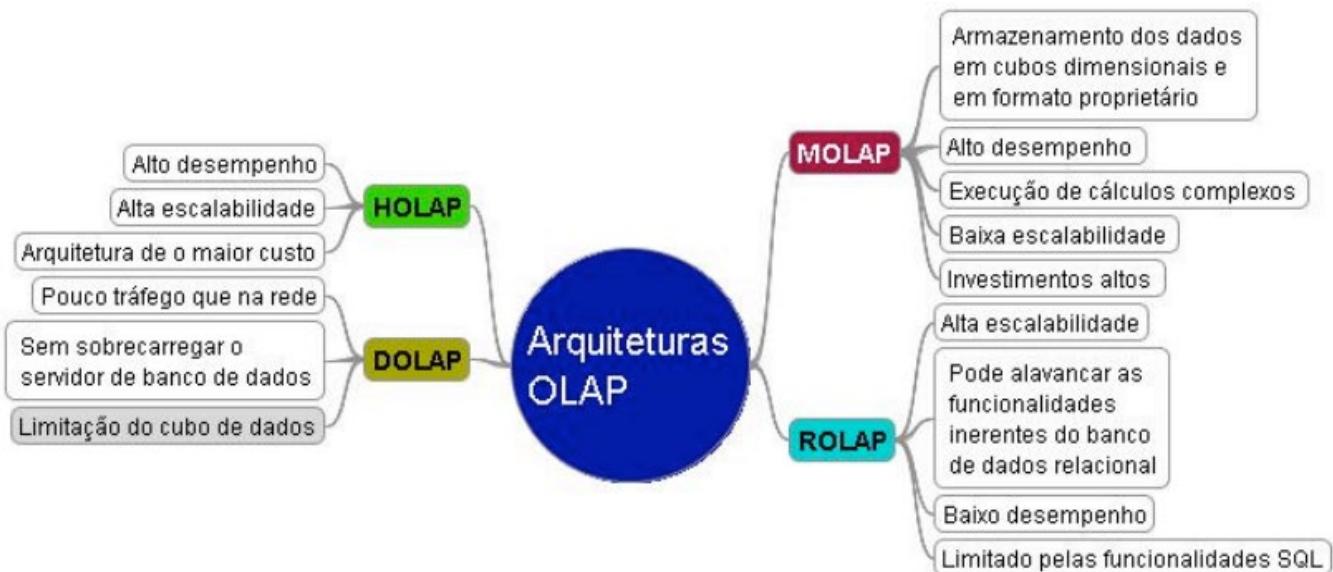
Fonte: Online Analytical Processing (OLAP) (2009).

- DOLAP (OLAP para Desktop): é a ferramenta para o usuário para que possa ter cópia da base multidimensional de dados ou um subconjunto dela, para acesso local na máquina.

Existem outras estruturas de armazenamento menos utilizadas. Na realidade, são uma versão das que foram apresentadas, principalmente, versões ajustadas para uso em desktop.

É possível, de forma resumida, apresentar as características dessas principais arquiteturas de dados conforme o que está apresentado na Figura 5.

Figura 6 – Característica das principais arquiteturas de dados



Fonte: Rodrigues et al. (2012).

Ainda, em se tratando de modelagem multidimensional, alguns conceitos característicos precisam ser definidos para um melhor entendimento do conceito de modelagem multidimensional: (1) medida; (2) fato; (3) dimensão e (4) hierarquia, os quais serão brevemente descritos a seguir.

- Medida: característica numérica que representa a mensuração de alguma informação, como porcentagens e quantidades.
- Dimensão: são as formas de visualização dos dados, de maneira hierarquizada (período, tipo, etc.).
- Hierarquia: classificação dos membros das dimensões (por exemplo, o período de tempo pode ser classificado em mês, dia, hora etc.).

ASSIMILE

Além de ROLAP, MOLAP e HOLAP, existem outras arquiteturas de dados, como, por exemplo, o JOLAP e o WOLAP. O JOLAP é fruto do esforço da Java Community



Process (JCP) em projetar uma API (*Application Programming Interface*) Java para servidores e para as aplicações dos sistemas OLAP. Já a WOLAP é arquitetura baseada em uso de sistemas OLAP em navegadores Web.

► 4. Visualização de dados com ferramentas OLAP

Segundo Thomsen (2002, p. 215, tradução nossa), “a visualização de dados é um assunto complexo e multifacetado”. Ainda segundo o mesmo autor, em se tratando de tecnologia, pode-se dizer que visualização de dados inclui aceleradores gráficos, bibliotecas gráficas, controladores gráficos, ambientes de desenvolvimento gráfico e aplicações gráficas, de forma geral. Quando se trata do usuário, pode-se dizer que visualização de dados inclui percepção visual, uso de algoritmos de renderização e representação e análise exploratória de dados.

Gráficos básicos, como os gráficos de linha e gráficos de barras, mostram no máximo duas ou três dimensões de informação. No entanto, muitas vezes, os usuários desejam trabalhar com conjuntos de dados OLAP compostos de mais de três dimensões. Por exemplo, um conjunto de dados de gerenciamento de campanhas pode ser composto pelas seguintes dimensões: produto, hora, canal, representante de vendas e tipo de promoção. Isso levanta a questão sobre a melhor maneira de visualizar mais de três dimensões dos dados.

Laudon e Laudon (2007. p.151 apud Rodrigues et. al., 2012, p. 2) afirmam que “o OLAP possibilita aos usuários obterem “respostas *online* a questões específicas”, de maneira rápida e eficiente, ainda que os repositórios sejam muito grandes ou que a análise leve em conta períodos longos.

Embora os dados empresariais típicos possam ter seis ou mais dimensões, é sempre possível obter uma parte ou um subconjunto dos dados e usar técnicas básicas de visualização que utilizam dimensões inferiores a três. Por exemplo, como mostra a Figura 5, um gráfico de linha básico, com duas dimensões, pode representar o preço de um ativo financeiro por tempo para um conjunto de dados de cinco dimensões que consiste em empresa, horários, produtos, cenários e a variável preço.

Figura 7 – Preço diário de um ativo financeiro de (1) empresas de papel para (2) homens investidores arrojados em (3) 2002 (4) na bolsa de valores de São Paulo



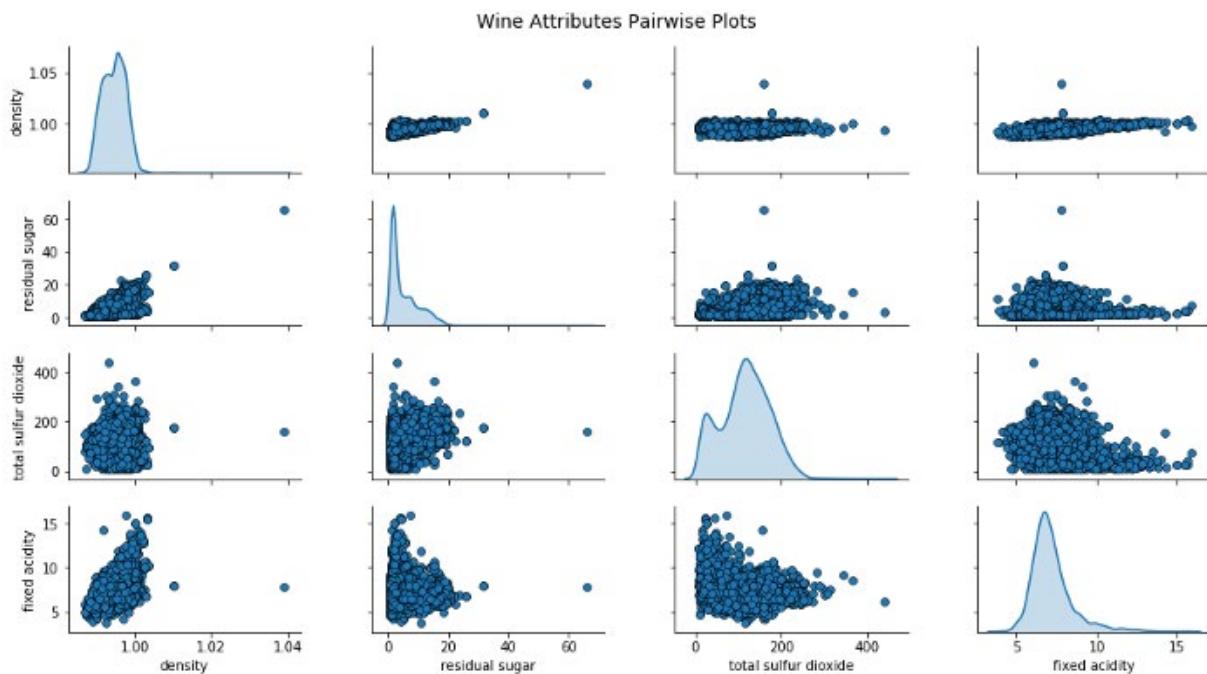
Fonte: Portal Action (2019).

Embora o padrão para os preços dos ativos na Bolsa de São Paulo sejam facilmente visualizados e satisfaçam as necessidades de um

consultor das informações, quando seu único interesse é o desempenho do ativo de determinado empresa na Bolsa de Valores de São Paulo. As informações contidas na visualização básica representam um subconjunto muito pequeno dos dados e é análogo à busca de uma agulha em um palheiro, pois foi necessário inserir uma série de filtros para se chegar no resultado de interesse.

E se você estivesse interessado em saber como as mudanças nos preços do ativo de outras empresas e em algumas corretoras ocorreram ao longo do tempo? Você pode estar tentando identificar onde houve menor variação nas vendas. Para fazer isso, seria necessário visualizar uma porcentagem muito maior dos dados. E se você ainda quisesse ver imagens em dimensões mais baixas, seriam necessárias muitas visualizações para apresentar todo o detalhamento de informações de interesse. Seria necessário construir o que se chama de matriz de gráficos. A Figura 8 apresenta um exemplo de matriz de gráficos, cuja finalidade é apresentar dados multidimensionais de forma gráfica.

Figura 8 – Exemplo de uma matriz de gráficos



Fonte: Sarkar (2018).

Anzanello (2002, apud Rodrigues et. al., 2012) apresenta alguns tipos de operação das ferramentas OLAP, as quais são: (1) consultas ad-hoc; (2) slice and dice e; (3) drill down/up. Uma breve descrição sobre cada uma delas será apresentada a seguir.

- Consultas ad-hoc: são geradas pelo usuário final de acordo com o que busca para obter informações para tomada de decisão.
- Slice and dice: permite que o usuário altere a perspectiva de visualização. Permite a modificação da posição das informações para facilitar a extração de resultados de interesse.
- Drill down/up: permite obter informações em diversos níveis de detalhamento, como, por exemplo, o tempo em ano, semestre, mês, dia, etc.

O uso desses recursos de operação OLAP permite obter uma série de produtos que ajudam na tomada de decisão, a identificação de padrões, tendências, dentre outras informações importantes para o acompanhamento de processos e gerenciamento de fluxos.

As ferramentas OLAP, assim como a visualização de dados feita por elas, aceleram o processo de decisão e mantêm a sua qualidade. Cada vez mais as empresas, no processo de competitividade, armam-se de ferramentas como essas, no intuito de assumirem a posição pioneira em seus ramos de atividade.

TEORIA EM PRÁTICA



Suponha que uma empresa possua um grande banco de dados com dados sobre suas movimentações financeiras, informações de clientes e de fluxo de colaboradores. Para um bom andamento dos processos internos, a empresa deseja disponibilizar para diversos departamentos

internos os dados armazenados, para que possam fazer planejamentos internos e acompanhamento dos processos em andamento. No entanto, a disponibilização das informações brutas apresenta, de certa forma, desvantagens, pois as informações a que os departamentos têm acesso, nem sempre são as mesmas. Algumas informações estão disponibilizadas para uns departamentos e outras para outros departamentos, além do que, com acesso aos dados brutos, poderá ser que os usuários manipulem os dados de forma inadequada, chegando em resultados errôneos.

Portanto, a empresa passou a investir em ferramentas de processamento *online* para apresentar aos departamentos os dados já com determinado nível de tratamento analítico para que sejam utilizados em seus planejamentos e acompanhamentos de processo.

Você faz parte da equipe responsável por apresentar as melhorias necessárias para o desenvolvimento dos processos de trabalho da empresa. O que você faria? E, por onde começaria?

VERIFICAÇÃO DE LEITURA

- 
1. O conceito de OLAP é amplo e se refere a um ambiente composto por diversas ferramentas tecnológicas que auxiliam no armazenamento e na exibição de manipulações de dados. OLAP, que é um acrônimo do inglês *online analytical process*, também é conhecido em

português como processo analítico *online*, mas ainda tem outra tradução. Qual é esta tradução?

Assinale a alternativa CORRETA.

- a. Processo de visualização de dados.
 - b. Sistema de informações multidimensionais.
 - c. Sistema de visualização de dados.
 - d. Processo *offline* analítico.
 - e. Sistema *offline* analítico.
2. Dada a sua amplitude de definição, o OLAP pode ser dividido em, segundo Thomsen (2002), conceito OLAP, linguagem OLAP e produtos OLAP. Quando se fala em produtos OLAP, estamos nos referindo a quais elementos?

Assinale a alternativa CORRETA.

- a. Computadores.
 - b. Computadores e métodos.
 - c. Compiladores e métodos de armazenamento.
 - d. Métodos.
 - e. Tipos de armazenamentos.
3. O conceito de OLAP, elaborado como um conjunto de programas computacionais, permite ao usuário realizar que tipo de análise?

Assinale a alternativa CORRETA.

- a. Multidimensional.
- b. Unidimensional.

- c. Multivariável.
- d. Univariável.
- e. Descritiva.

► Referências Bibliográficas

- Analytical Processing (OLAP). 2009. Disponível em: <http://thebusinessintelligence.blogspot.com/2009/12/online-analytical-processing-olap.html>. Acesso em: 23 jul. 2019.
- PORTAL ACTION. 2019. Disponível em: <http://www.portalaction.com.br/estatistica-basica/33-grafico-de-linhas>. Acesso em: 07 ago. 2019.
- RIBEIRO, V. **O que é OLAP?** 2013. Disponível em: <https://vivaneribeiro1.wordpress.com/2011/07/12/o-que-e-olap/>. Acesso em: 07 ago. 2019.
- ROCCO, C.V. **Implantação de um ambiente de business intelligence como apoio a decisões empresariais.** 2009. 31f. Monografia (Bacharelado em engenharia de computação) – Curso de Engenharia de computação da Universidade São Francisco, Campus de Itatiba. Disponível em: <http://lyceumonline.usf.edu.br/salavirtual/documentos/1720.pdf>. Acesso em: 23 jul. 2019.
- RODRIGUES, C.H.M.; ALMEIDA, C.C.O.; ROCHA, E.D.; COSTA, E.J.A. OLAP: uma perspectiva estratégica de análise de dados. **Revista Clique**, v. 1, n. 1, ago. 2012. Disponível em: <http://webcache.googleusercontent.com/search?q=cache:ECFoJ3-fI4J:www.periodicos.unimontes.br/clique/article/download/67/37+&cd=4&hl=pt-BR&ct=clnk&gl=br>. Acesso em: 23 jul. 2019.
- SARKAR, D. **The art of effective visualization of multi-dimensional data: strategies for effective data visualization.** 2018. Disponível em: <https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57>. Acesso em: 07 ago. 2019.
- VIEIRA, E. **Tecnologia olap.** 2009. 38f. Trabalho de conclusão de curso (Graduação em Ciência da Computação). Instituto Municipal do Ensino Superior de Assis, Assis, 2009. Disponível em: <https://cepein.femanet.com.br/BDigital/arqTccs/0411150200.pdf>. Acesso em: 22 jul. 2019.
- THOMSEN, E. **OLAP solutions:** building multidimensional information systems. 2. Ed. New York: John Wiley & Sons, Inc. 2002.



► Gabarito

Questão 1 – Resposta: B.

Resolução: Os sistemas OLAP também são conhecidos em português como sistema de informações multidimensionais.

Feedback de reforço: Lembre-se do texto inicial de nossa aula. Releia se necessário a leitura fundamental.

Questão 2 – Resposta: C.

Resolução: Os produtos OLAP incluem compiladores e métodos de armazenamento e de acesso.

Feedback de reforço: Lembre-se da divisão que Thomsen (2009) faz sobre OLAP. Reveja na leitura fundamental.

Questão 3 – Resposta: A.

Resolução: O conceito OLAP como um conjunto de programas computacionais permite ao usuário usufruir de análises multidimensionais.

Feedback de reforço: Lembre-se do conceito de OLAP como um conjunto de softwares. Releia a leitura fundamental, se necessário.



Data discovery

Autor: Marcelo Tavares de Lima

► Objetivos

- Apresentar conceitos básicos de *data discovery*.
- Apresentar conceitos fundamentais associados a *data discovery*.
- Identificar contextos apropriados para aplicação de conceitos de *data discovery*.

1. Introdução

A descoberta de informações importantes, padrões e tendências que auxiliem na tomada de decisões através de processos de visualização de dados é um dos grandes desafios das empresas e dos negócios, em geral.

O processo de descoberta de dados, conhecido como *data discovery*, tem como principal objetivo encontrar padrões e anomalias para dar suporte às decisões das empresas e interessados, com intuito de sempre estarem à frente em termos de competitividade de mercado.

Neste texto serão abordados conceitos fundamentais sobre *data discovery* e, também, conceitos associados a este processo, o qual tem muita relação com os conceitos de *business intelligence* (BI).

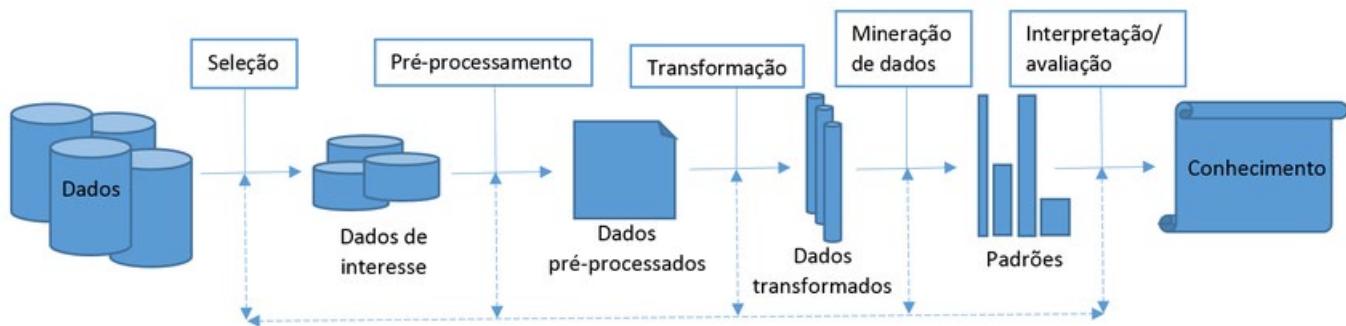
Desejamos bons estudos!

2. Conceitos básicos de *data discovery*

A descoberta de conhecimento através de bases de dados (KDD, do inglês *knowledge Discovery in Databases*), de forma resumida, pode ser considerada como o processo de identificação de padrões novos, padrões válidos ou que façam algum sentido, que possam ser úteis e, principalmente, comprehensíveis e interpretáveis.

A descoberta do conhecimento em bases de dados tem como principal objetivo a extração de conhecimento inteligível e utilizável para o apoio à tomada de decisões, e encontrar padrões e tendências. O processo de descoberta do conhecimento através dos dados é composto por algumas etapas, conforme ilustradas na Figura 1.

Figura 1 – Processo de descoberta de conhecimento em bancos de dados



Fonte: Freitas, Moura e Silva (2015).

É claro que se trata de um ciclo, onde pode ser que algumas etapas necessitem de nova realização por conta de uma série de situações que influenciem no desenvolvimento do processo.

A descoberta de dados, ou mais comumente conhecida como *data discovery*, é uma maneira de permitir que os envolvidos em algum processo de trabalho e/ou negócios possam obter as informações necessárias para realizar seus trabalhos de forma confiável, intuitiva e de alta disponibilidade. Esse seria um conceito amplo.

Segundo Ryan (2016, p. 27, tradução nossa)

data discovery-ou, considerando seu apelido, a *information discovery* (ID)-não é exatamente novo. De fato, o passado viu a descoberta de informações serem referidas como, em tom de brincadeira, “uma das tecnologias mais legais e assustadoras em BI”.

Em termos práticos, existem muitas ferramentas que colocam em prática o conceito de *data discovery*, como o Tableau, Qlik Sense, Python e R, dentre outros programas que implementam técnicas estatísticas e algoritmos matemáticos que auxiliam na descoberta de informações para a aquisição de conhecimento para melhorar as tomadas de decisões e análises especializadas.

Os dados trabalhados no contexto do *data discovery* têm origem diversa, diferentes instantes de coleta e em lugares distintos. Essa diversidade gera a necessidade de um esforço para consolidação e agrupamento para a geração de informação confiável e interpretável para o processo, isso até faz lembrar o termo de armazenamento de dados ou *data warehouse*.

O entendimento do negócio e do contexto onde os dados estão inseridos é de fundamental importância para o entendimento dos mesmos. Dada a diversidade e heterogeneidade dos dados, os esforços para a consistência dos mesmos são de extrema importância para que, em seguida, possam vir a ser trabalhados para a busca de conhecimento útil.

Enfim, o processo de *data discovery*, em termos práticos, significa detectar padrões a partir de dados com o suporte humano ou através de sistemas de inteligência artificial. Como resultado, os dados tratados são geralmente apresentados na forma de gráficos e visualizações diversas.

Alguns especialistas afirmam que existem muitas semelhanças entre *data discovery* e *data mining*, já outros afirmam que são essencialmente processos diferentes. O que se pode dizer sobre ambos é que possuem características de utilizarem grandes volumes de dados para detectar informações relevantes.

Também há muitas semelhanças entre *business intelligence* e *data discovery*, mas alguns especialistas afirmam que o *business intelligence* é a fundação das organizações, enquanto que o *data discovery* seria um complemento. Vale lembrar que o conceito de *business intelligence* (BI) é o processo de abordagem tradicional para suporte na tomada das melhores decisões baseadas em dados. De forma tradicional, o BI tem relação direta com o desenvolvimento de um depósito de dados ou um *data warehouse*, que é usado para elaboração de relatórios e de análises.

Os dados de um BI, em geral, são dados internos oriundos de programas de sistemas integrados de gestão empresarial (ERP), gestão de relacionamento com o cliente (CRM) ou outros sistemas importantes para a organização.

Especialistas, como Ryan (2016) e Matthew (2019), afirmam que o *data discovery* não requer *data warehouse*, além de fazer integração entre dados de origens distintas, inclusive, é claro, com ERP e CRM. Outro diferencial do conceito de *data discovery*, segundo especialistas da área, é que o seu processo permite identificar perguntas sem respostas, ou seja, é possível identificar itens ou elementos que não seriam possíveis de ver com outra metodologia. Isso ocorre, principalmente, com a visualização de dados, ou seja, o *data discovery*, onde é possível identificar padrões e tendências mais facilmente e, a partir disto, levantar hipóteses e questionamentos.

Ryan (2016, p. 30, tradução nossa) afirma que “a diferença fundamental entre o BI e o *data discovery* é simples: um começa com uma pré-definição e expectativa dos dados, enquanto o outro termina com uma nova definição derivada de novos *insights* sobre os dados”.

Data discovery, que não deve ser confundido com exploração de dados, tem como objetivo inicial alcançar os objetivos dos negócios, mas permite que não saibamos quais métricas ou quais dados precisamos para atingir a meta. Esta é a natureza investigativa do *data discovery*, explorando, visualizando, separando e torturando os dados em um processo iterativo para descobrir relações, padrões e tendências nos dados.

Pode-se até já saber o contexto dos dados, no entanto, o objetivo é construir novos modelos com o *data discovery* para descobrir relações desconhecidas e, então, encontrar maneiras de como essa informação poderá fornecer valor para o negócio, enquanto a empresa encontra-se em evolução.



PARA SABER MAIS

Discovery versus exploração de dados. A exploração de dados é um processo de investigar, examinar ou procurar novas informações. Dentro do processo de descoberta (*discovery*), a exploração é a jornada inicial. Explorar é um precursor da descoberta. A descoberta (*discovery*) em si é o jogo final da exploração: é dar a conhecer ou expor algo que era desconhecido.

► 3. Formas tradicionais de *data discovery*

De início, tanto as planilhas de dados, como MS Excel, quanto as visualizações básicas, como gráficos básicos (barras, linhas e setores), são consideradas formas tradicionais de *data discovery*.

3.1 Planilhas

As planilhas, como o MS Excel, continuam sendo a ferramenta mais popular de *business analytics*¹ para trabalhar com dados, em parte devido à sua ampla existência e disponibilidade e devido à familiaridade do usuário, dada facilidade de sua utilização.

Uma planilha de dados pode ser uma poderosa ferramenta analítica nas mãos de um usuário experiente. Com as primeiras planilhas, como a VisiCalc e a Lotus 1-2-3, descobriu-se um poderoso paradigma para os humanos organizarem, calcularem e analisarem dados que têm resistido ao teste do tempo, embora algumas empresas não as utilizam mais para trabalhos analíticos. Em contrapartida, o Microsoft Excel, em suas últimas versões, conta com mais de

¹ Inteligência de negócios com abordagem em análise de dados.

um milhão de linhas (1.048.576) e mais de 16.000 colunas (16.384) de dados, aumentando sua capacidade de armazenamento e de tratamento de dados

3.2 Visualizações básicas

Visualizações básicas, como gráficos de barras, linhas ou de setores (incluindo aqueles incorporados em *dashboards*) gerados ou não no MS Excel, fornecem informações simples e diretas por representações visuais de dados que permitem aos analistas encontrar *insights* que podem não ser tão facilmente percebidos em um formato de texto simples.

Não é uma tarefa fácil afirmar exatamente o que seja uma visualização básica de dados, mas talvez seja uma descrição simples em dizer que as visualizações básicas são um meio eficaz de descrever, explorar ou resumir os dados, porque o uso de uma imagem pode simplificar informações complexas e ajudam a destacar ou descobrir padrões e tendências. Elas também podem ajudar a apresentar grandes quantidades de dados e podem facilmente ser utilizadas para apresentar conjuntos de dados menores também.

Entretanto, as visualizações básicas ficam aquém em relação aos métodos de visualização mais avançados em muitos aspectos. Além disso, elas não permitem o uso dinâmico de dados e não oferecem habilidades para consultas dinâmicas, nem mecanismos de monitoramento em tempo real.

4. Formas avançadas de *data discovery*

A evolução das formas tradicionais de *data discovery* levou a novas e avançadas formas de descoberta que permitem pesquisar e visualizar

vários tipos de dados dentro de um ambiente ou processo de gestão para tomada de decisão.

4.1 Modo de pesquisa multifacetado

Um *data discovery* multifacetado (ou “modo de busca”) permite aos analistas minerar dados para obtenção de *insights* sem discriminar entre dados estruturados e não estruturados. Podem acessar dados em documentos, e-mails, imagens, redes sociais, etc., em um mecanismo de pesquisa como o Google, Yahoo! ou Bing, com a capacidade de iteragir conforme suas necessidades, além de poder detalhar a fundo os dados disponíveis para descobrir novos *insights*. O computador IBM Watson, por exemplo, é uma ferramenta de *data discovery* capaz de responder perguntas levantadas no dia a dia dos negócios, de forma iterativa e dinâmica.

4.2 Visualizações avançadas

As visualizações avançadas são tudo o que as visualizações básicas de dados não são, são aquelas obtidas com ferramentas mais especializadas, com arquiteturas de dados mais estruturadas. Elas são ferramentas para *visual discovery*, que permitem aos analistas utilizar *big data* para descobrir *insights* de maneira totalmente nova. As visualizações avançadas podem complementar formas tradicionais de descoberta para dar oportunidade de comparação entre vários tipos de *data discovery*, para potencializar descobertas de *insights* ou para ter uma visão mais completa dos dados.

Com visualizações avançadas, os analistas podem identificar *clusters* ou agregados diversos; também podem analisar dados para procurar correlações ou preditores para descobrir novos modelos analíticos.

As visualizações avançadas podem ser obtidas por processos interativo, compostas por animações e, em alguns casos, podem fornecer análise de dados em tempo real. Além disso, as visualizações avançadas são multidimensionais, proporcionando oportunidades de descoberta visual ideais, o que dificilmente ocorrerá com visualizações básicas.

Os tipos de visualização de dados básicas podem ser otimizadas para se tornarem visualizações de dados avançadas, com a inclusão de dimensões e artifícios para torná-las mais poderosas.

A inclusão de artifícios visuais, como ícones inteligentes e graduações de cores em mapas, por exemplo, são recursos utilizados em *visual discovery* avançado, que agrupa melhores práticas em ciências cognitivas e no *design visual*.

Lembre-se, as visualizações avançadas não são simplesmente uma função de como os dados podem ser visualizados, mas são medidas de como os dados podem ser dinâmicos, interativos e funcionais.

ASSIMILE



A trajetória do *data discovery*. Os sistemas de *data discovery* estão começando a cumprir promessas feitas até então não cumpridas. O avanço das ferramentas analíticas está proporcionando a cada dia que se avança a possibilidade de maior exploração de dados. É claro que há ainda áreas a melhorar para fechar lacunas e encontrar equilíbrio em recursos e funcionalidades.

5. Ferramentas de *data discovery*

A maioria das pessoas de negócios não são programadores ETL (do inglês, *extract, transform, load*), ou seja, não detêm o conhecimento necessário para realizar integração de dados; não conhecem a linguagem SQL ou qualquer outra linguagem e, muitos, sequer têm alguma formação para realizar análise de dados.

Em vez de considerarmos estas características como uma fraqueza desses profissionais, devemos considerar tal situação como uma oportunidade. Os profissionais dos negócios precisam de ferramentas que trabalhem de acordo com as suas necessidades sem precisar ter que escrever códigos de programação ou ter que realizar alguma atividade específica dos profissionais de tecnologia da informação (TI).

Basicamente, os profissionais de negócios buscam ferramentas de *data discovery*, que sejam autossuficientes para duas coisas. Primeira, precisam de ferramentas ágeis, iterativas, que permitam interação entre elas e os dados. Em segundo lugar, juntamente com as melhores práticas, a descoberta guiada e outros mecanismos para preservar a integridade do processo de *data discovery*, afastar situações de risco para os negócios.

Pelas razões apresentadas, essas ferramentas de *data discovery* precisam ser visuais e convidativas para o *visual discovery* e *storytelling*², no intuito de promover o pensamento e a comunicação convincente de *insights*.

Juntas, as duas razões apresentadas são requisitos básicos para a autossuficiência: o melhor do BI, juntamente com a melhor visualização de dados. Cumulativamente, elas dão aos usuários a capacidade de integrar e abstrair dados, visualizá-los, analisá-los

² Descrição de situações e eventos através de recursos visuais.

para obter *insights*, compartilhar descobertas e fazer tudo de novo, se necessário.

É óbvio que as ferramentas de *data discovery* não se bastam para dar autonomia aos seus usuários. Junto com essas ferramentas, os usuários precisam de bons mecanismos tecnológicos para trabalhar com dados e não apenas um conjunto de dados.

► 6. Ambientes de *data discovery*

O processo de descoberta de dados em negócios precisa ser comandado por seus usuários, ou seja, as pessoas envolvidas nos processos de negócios. Tais usuários são os guias com o conhecimento necessário para identificar oportunidades e saber o que estão procurando. Podem até não saber o que procuram, no entanto, quando surgir alguma pista, poderão fazer essa identificação.

A descoberta de dados requer um salto de fé. Essa mentalidade otimista e voltada para o futuro é uma das principais distinções entre BI e *data discovery*. Embora ambos se concentrem em extrair informações dos dados, BI e *data discovery* se diferenciam, basicamente, em: BI os usuários entram nos dados com ideias prontas, enquanto que em *data discovery*, utilizam os dados para obter *insights*.

Por fim, torna-se praticamente uma obrigação deixar os usuários autossuficientes. A equipe de TI precisa instalar ambientes e sistemas que permitam tal situação. Situações de ambientes favoráveis ao uso autossuficientes de ferramentas de BI e *data discovery* estimulam o processo de criação de ideias e de buscas por *insights* para a tomada de decisão apropriada.

7. O cientista de dados

Já dizia o economista chefe da Google, em 2016, que “a profissão mais sexy nos próximos dez anos será a de estatístico”. No entanto, o artigo de Davenport e Patil (2012) afirmava algo mais amplo, inclusive, é o título do artigo que traz a afirmação dizendo que cientista de dados é o trabalho mais sexy do século 21.

Amaral (2016, n.p.) descreve o cientista de dados como “alguém com conhecimento técnico vertical em estatística, NoSQL, *cloud computing*, mineração de dados (*data mining*), álgebra relacional, modelagem multidimensional, MapReduce, virtualização, entre outros”. No entanto, Amaral (2016) ainda descreve que há uma exagerada carga de conhecimento atribuída a este profissional que, em termos comparativos, só poderia ser feito com uma máquina superpotente, como o supercomputador Watson, da IBM.

O que se percebe, comparativamente ao surgimento do termo dado ao profissional multidisciplinar cientista de dados, é que o cenário tem apontado para um perfil um tanto diferente do definido inicialmente.

O cientista de dados pode estar habilitado para atuar na instalação dos sistemas e ambientes de *data warehouse*, no gerenciamento de bancos de dados e produção de informações e atuação em *data discovery*. É amplo o seu campo de atuação.

Surgiu, também, ao longo do avanço tecnológico, tanto dos sistemas de gerenciamento de dados quanto das ferramentas de visualização, o profissional conhecido como analista de visualização de dados. Uma descrição sobre tal profissional é dada por Ryan (2016), apresentada de forma resumida no item a seguir.

8. O analista de visualização de dados

Com maior ênfase na visualização de dados do que o cientista de dados, o *data visualization analyst* tem atividades focadas na elaboração de insumos visuais que sejam, no mínimo, animadas, interativas e de alta performance, o papel do analista de visualização de dados é tomar grandes quantidades de informações e condensá-las nos gráficos, mapas ou infográficos orientados a dados para exibir informações substanciais.

É um profissional com papel mais técnico em habilidades e com a necessidade de ter uma profunda compreensão de *analytics* e das ferramentas analíticas para processar dados de várias fontes (assim como extrair, transformar e carregar os dados nos diversos relatórios gerenciais) e torná-los significativos dentro do contexto dos objetivos. No Brasil, essas atividades desse profissional ainda estão associadas ao cientista de dados. Este profissional precisará de habilidades de linguagem de programação bastante sofisticadas para construir (desenvolver) ou, no mínimo, usar ferramentas de usuário orientadas para autoatendimento para projetar visualização de dados adequadas.

Enfim, na prática, o que o mercado precisa, de fato, é de profissionais multidisciplinares, os quais deverão possuir conhecimentos em tecnologia da informação, modelagem e conceitos de negócios. Este, com um conhecimento amplo, poderá atuar como líder e se cercar de profissionais específicos de cada área.

Amaral (2016) apresenta um quadro comparativo entre o que se deseja ou almeja de um profissional das ciências de dados com o profissional, normalmente, disponível no mercado. O Quadro 1 a seguir faz uma replicação do apresentado.

Quadro 1 – O Cientista de dados disponível e idealizado

Profissional de mercado	Profissional idealizado
Conhecimento multidisciplinar	Especialista em todas as áreas
Gerência de projetos	Foco em conhecimento técnico
Liderança	Trabalha sozinho
Equipe de especialistas	Especialista em todas as áreas

Fonte: Amaral (2016).

O processo de *data discovery* é um conceito amplo que está associado a um conjunto de conceitos e termos da tecnologia da informação (TI), de estatística, dentre outros.

Existem inúmeras ferramentas no mercado que atuam e disponibilizam a possibilidade de realizar atividades de gestão através da visualização de dados, o que tem aumentado cada vez mais a disputa pelo espaço nesta área, tanto de quem produz tecnologias quanto por quem é habilitado em manipulá-las.

TEORIA EM PRÁTICA



Inteligência Artificial. Existe uma série de matérias nos diversos sites de notícias que afirmam que os robôs vão dominar o mundo. Eles afirmam ainda que, por conta disso, muitas profissões serão extintas e os profissionais que ocupam essas atividades deverão se reinventar para poder continuar no mercado de trabalho. Trindade (2018) apresenta uma série de números e informações, inclusive as profissões mais ameaçadas de extinção. A matéria mostra ainda o percentual de chances de substituição de seres humanos por robôs em algumas áreas, como, por exemplo, na preparação de comidas, a matéria afirma que existe uma

chance de 64% de os robôs substituírem os seres humanos. Matérias como essa surgem cada vez mais na mídia. No entanto, não são motivos de medo ou pavor. O que elas trazem serve como estímulo para que os profissionais se reciclem e se reinventem de forma contínua, senão ficam para trás na corrida que sempre existiu pela disputa de uma vaga no mercado de trabalho.

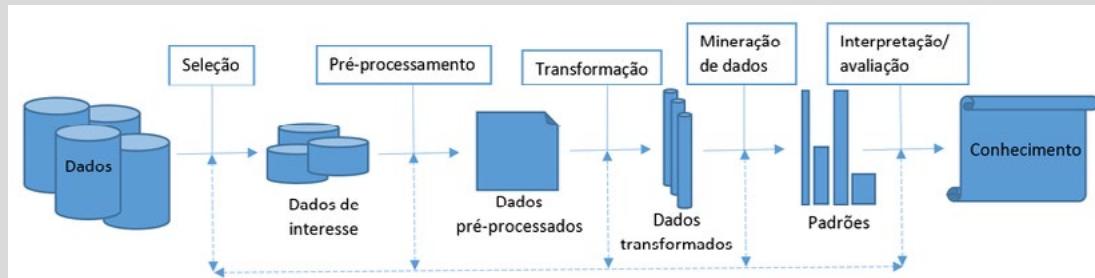
Agora que você conhece o conceito de inteligência artificial, avalie a sua profissão e responda às seguintes perguntas: Você acha que a sua profissão corre risco de extinção com a inteligência artificial? Por quê?

VERIFICAÇÃO DE LEITURA



1. A descoberta de dados ou o *data discovery* é um conceito que abrange uma série de etapas de atividades. Segundo a Figura 1 deste material (apresentada novamente a seguir), qual a etapa que se segue à primeira etapa e que se refere à disponibilização de dados?

Figura 1 – Processo de descoberta de conhecimento em bancos de dados



Fonte: Freitas, Moura e Silva (2015).

Assinale a alternativa CORRETA.

- a. Disponibilização de dados.
- b. Dados de interesse.
- c. Dados transformados.
- d. Produção de conhecimento.
- e. Identificação de padrão.

2. Segundo Ryan (2016), a nomenclatura *data discovery* possui um apelido. Que apelido é esse?

Assinale a alternativa CORRETA.

- a. Data *Information*.
- b. Data Science.
- c. Information Discovery.
- d. Information Science.
- e. Data visualization.

3. O conceito de *data discovery* tem associação com o uso de grandes bases de dados. Outro termo técnico que também se utiliza de grandes bases de dados faz com que especialistas façam associação com *data discovery*. A qual termo estamos nos referindo?

Assinale a alternativa CORRETA.

- a. *Data discovery*.
- b. Inteligência artificial.
- c. *Data visualization*.
- d. *Data mining*.
- e. *Machine learning*.

Referências Bibliográficas

AMARAL, Fernando. **Introdução a ciência de dados:** mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016. KINDLE. Não paginado.

Davenport, T.H., Patil, D.J. Data scientist: the sexiest job of the 21st century. **Harvard Bus. Rev.**, 90 (10), 70–76, 2012. Disponível em: <http://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?vid=3&sid=8165203e-e4df-4dea-b498-8e70cf216046%40pdc-v-sessmgr03>. Acesso em: 25 jul. 2019.

Freitas, N.; Moura, C.; Silva, M. (2015). Sistema multiagente para mineração de imagens de satélite. **Anais XVII Simpósio Brasileiro de Sensoriamento Remoto – SBSR**. João Pessoa-PB, 25 a 29 de abril de 2015. Disponível em: https://www.researchgate.net/publication/283716349_Sistema_multiagente_para_mineracao_de_imagens_de_satelite. Acesso em: 25 jul. 2019.

MATTHEW, N. **Why visual analytics?** Tableau. 2019. Disponível em: https://www.tableau.com/sites/default/files/whitepapers/752750_core_why_visual_analytics_whitepaper_0.pdf. Acesso em: 24 jul. 2019.

RYAN, L. **The visual imperative:** creating a visual culture of *data discovery*. Cambridge: Elsevier, 2016.

TRINDADE, R. **A máquina no lugar do homem: A inteligência artificial eliminará empregos, mas novas profissões surgirão; qualificação será fundamental.** 17/08/2018. Disponível em: <https://www.uol/tecnologia/especiais/inteligencia-artificial-vai-acabar-com-empregos-.htm>. Acesso em: 19 ago. 2019.

Gabarito

Questão 1 – Resposta: B.

Resolução: Segundo o que se apresenta na Figura 1, que trata do processo de descoberta de dados de forma abrangente, seguindo a disponibilização de dados, tem-se a seleção dos dados de interesse.

Feedback de reforço: Lembre-se das etapas do processo de descoberta do conhecimento em bancos de dados, que passa pela seleção dos dados de interesse, pré-processamento, transformação, mineração e identificação de padrões, interpretação e avaliação e, enfim, o conhecimento.

Questão 2 – Resposta: C.

Resolução: Segundo Ryan (2016), o apelido dado ao *data discovery* é *information discovery*.

Feedback de reforço: Lembre-se que, no passado, viu-se o *data discovery* serem referidas como, em tom de brincadeira, “uma das tecnologias mais legais e assustadoras em BI”.

Questão 3 – Resposta: D.

Resolução: Por se relacionarem com grandes bases de dados, especialistas fazem muita associação entre *data discovery* e *data mining*.

Feedback de reforço: Lembre-se de que não há um consenso entre os especialistas da área sobre *data discovery* e *data mining*. No entanto, em um quesito apenas existe um consenso.



Outras ferramentas para visualização de dados (Chart.js, Leaflet, Datawrapper, Dygraphs, Highcharts, Google Charts, Polymaps e Weka)

Autor: Marcelo Tavares de Lima

► Objetivos

- Apresentar algumas ferramentas de visualização de dados.
- Descrever as especificidades de algumas ferramentas de visualização de dados.
- Demonstrar exemplos de aplicação de algumas ferramentas de visualização de dados.

1. Introdução

Muitas são as ferramentas de visualização de dados. Cada uma delas apresenta características distintas uma das outras. Como, por exemplo, algumas exigem conhecimento de linguagem de programação. Outras, são mais intuitivas e não têm essa exigência.

A intenção deste texto é apresentar algumas ferramentas de visualização de dados, sem a intenção de esgotar o assunto, pois é praticamente impossível e, também, descrever suas especificidades e utilidades para cada situação de trabalho e estudo, de acordo com a necessidade do usuário. Pretende-se com isso, apresentar a você, aluno, as principais ferramentas de visualização de dados do mercado, tornando-o apto a pesquisar e analisar qual delas se aplica melhor às suas necessidades.

2. Chart.js

É uma das bibliotecas de JavaScript utilizada para visualização de dados na Web. JavaScript é uma das linguagens de programação para Web muito utilizada para gerar visualização de dados (FLANAGAN, 2013). Boa parte dos sites modernos fazem uso dessa linguagem. O JavaScript é parte de uma tríade de tecnologias bastante conhecida por quem é desenvolvedor Web: HTML, que especifica o conteúdo dos sites; CSS, responsável por especificar a apresentação de páginas Web; e JavaScript, responsável por especificar o comportamento das páginas.

Segundo Machado Neto (2013, p. 40), “Chart.js é uma biblioteca desenvolvida por Nick Dowine sob a linguagem JavaScript, renderiza os gráficos utilizando o elemento canvas do HTML5”. Nesta biblioteca existem seis tipos de gráficos disponíveis: linhas, barras verticais, radar, área polar, pizza ou setores e rosca. A biblioteca Chart.js não exige dependências para sua execução, o que é caracterizado com uma

vantage em relação a outras bibliotecas JavaScript, além de ser muito leve para se trabalhar. Além dos gráficos apresentados na Figura 1, já existem implementados, também, gráficos de bolhas e de dispersão. A Figura 1 mostra um exemplo de cada um desses tipos disponíveis na biblioteca Chart.js.

Figura 1 – Tipos de gráficos disponíveis em Chart.js



Fonte: Machado Neto (2013).

Segundo Cintra (2018), a biblioteca Chart.js tem como principais características ser de código aberto (*open source*), disponibilizar um conjunto de gráficos que permitem a inserção de animação, estar baseada em HTML5 Canvas, ser responsiva, etc.

Para ser utilizada, a biblioteca Chart.js precisa que sejam seguidos alguns passos, conforme apresentado por Cintra (2019): 1) realizar o *download* pelo GITHUB ou CDN no endereço <https://www.chartjs.org/docs/latest/getting-started/installation.html>; 2) Fazer a definição de um elemento HTML5 Canvas; 3) Como opção, preparar estilos CSS (linguagem de programação mais robusta que o a HTML para estilizar páginas da Web); 4) Determinar o conjunto de gráficos a ser exibido;

e 5) Criar o script ou programação para renderizar (produzir) o(s) gráfico(s) que se deseja.

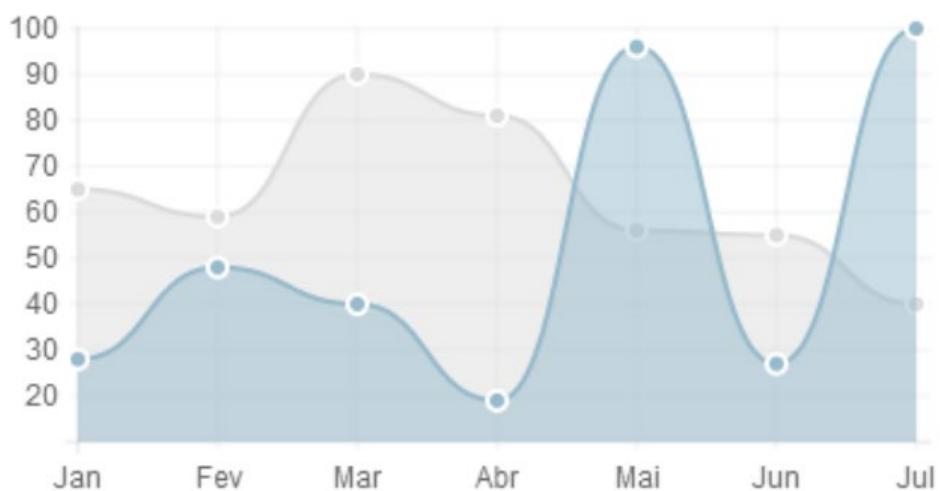
O Quadro 1 apresenta um script para geração de uma visualização com gráfico de linha e seu resultado, o gráfico em si, na Figura 2, ambos extraídos de Machado Neto (2013).

Quadro 1 – Código para renderizar um gráfico de linhas

```
// Opções para o gráfico
var options = {
    // Mostrar pontos no gráfico
    pointDot: true,
    // Animar o gráfico quando carregá-lo pela primeira vez
    animation: true
};
// Dados para plotagem
var data = {
    // Etiquetas para o eixo X para seus respectivos valores de pontos
    labels : ["Jan", "Fev", "Mar", "Abr", "Mai", "Jun", "Jul"],
    datasets : [
        {
            // Cor para a primeira linha
            fillColor : "rgba(220,220,220,0.5)",
            strokeColor : "rgba(220,220,220,1)",
            pointColor : "rgba(220,220,220,1)",
            // Cor do contorno dos pontos
            pointStrokeColor : "#fff",
            // Dados para plotagem
            data : [65,59,90,81,56,55,40]
        },
        {
            // Cor para a segunda linha
            fillColor : "rgba(151,187,205,0.5)",
            strokeColor : "rgba(151,187,205,1)",
            pointColor : "rgba(151,187,205,1)",
            // Cor do contorno dos pontos
            pointStrokeColor : "#fff",
            // Dados para plotagem
            data : [28,48,40,19,96,27,100]
        }
    ]
};
// Obtém o identificador do canvas e coloca o contexto como duas dimensões
var ctx = document.getElementById("chart").getContext("2d");
// A partir do canvas selecionado chama a função de plotagem de linha da biblioteca
// passando os dados e as opções selecionadas
var myNewChart = new Chart(ctx).Line(data,options);
```

Fonte: Machado Neto (2013).

Figura 2 – Gráfico de linhas produzido para visualização na Web, elaborado com a utilização de Chart.js



Fonte: Machado Neto (2013).

► 3. Leaflet.js

Biblioteca JavaScript, apropriada para a geração de mapas na internet e dispositivos móveis. Sua principal vantagem em comparação a um dos geradores de mapas mais conhecidos, o Google Maps, é que permite a troca entre fornecedores de mapas com facilidade, o que agrega flexibilidade à exibição de mapas de diversas fontes. Assim como a biblioteca Chart.js, também é de código aberto (*open source*) e tem código disponibilizado também no GitHub.

A página da biblioteca <https://leafletjs.com/> afirma que ela tem apenas 38KB de JavaScript, sendo considerada leve, simples e, por isso, de muito interesse da maioria dos desenvolvedores de mapas para desktops e dispositivos móveis.

É possível utilizar a biblioteca Leaflet dentro do RStudio. A vantagem de utilizar desta forma é que não há uma exigência de profundo conhecimento em linguagem JavaScript. Tran (2019) apresenta script em R para a geração de mapas e divulgação na Web.

O pacote Leaflet do R foi criado pelos desenvolvedores de linguagem R para o RStudio, com o intuito de integrar com a linguagem JavaScript. A seguir são apresentados comandos de programação em R para gerar mapas para a Web.

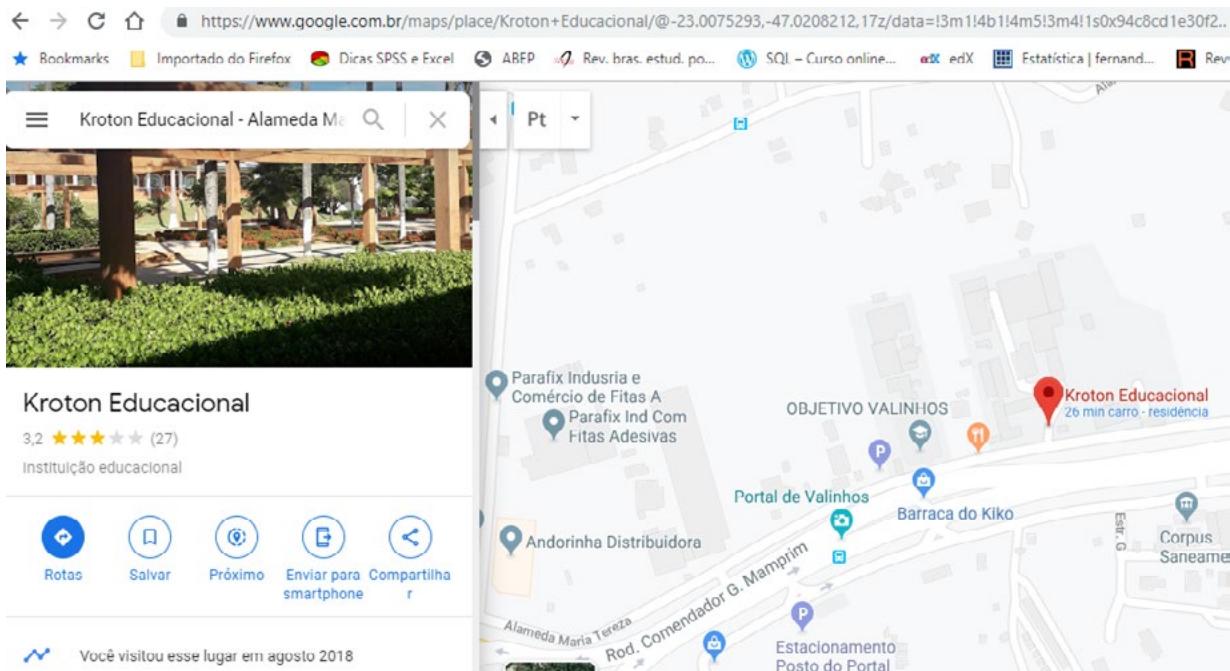
1. Necessário instalar as seguintes bibliotecas com os seguintes comandos.

```
library(leaflet)
```

```
library(dplyr)
```

2. Precisamos obter a latitude e a longitude em um mapa. Vamos ao Google Maps e buscar um endereço. Vamos selecionar o endereço do corporativo da Kroton, em Valinhos/SP, conforme mostra a Figura 3.

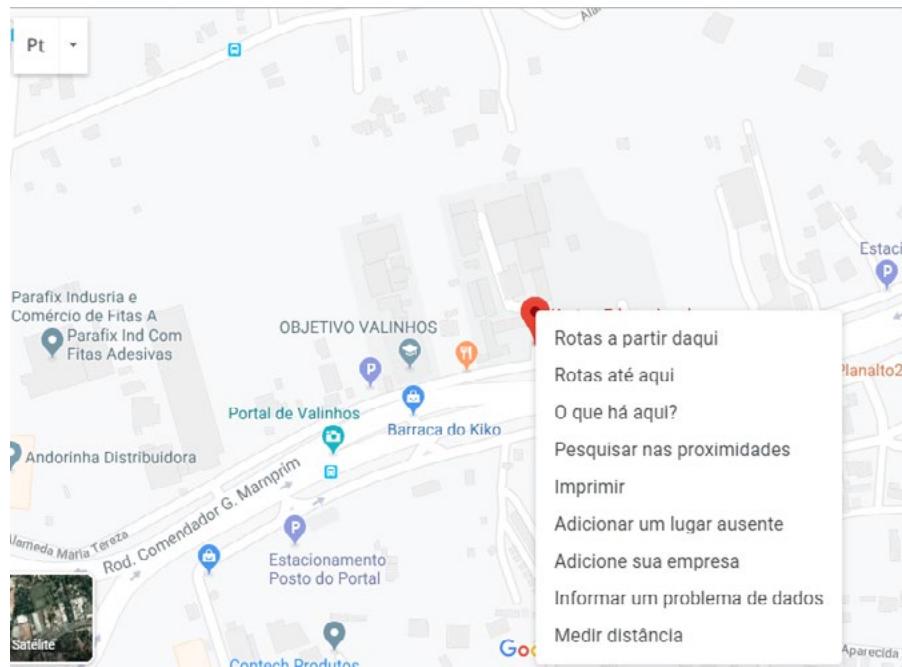
Figura 3 – Mapa do Google maps com seleção de um endereço específico



Fonte: Google maps.

3. Clique com o botão direito do mouse em cima do ícone vermelho que está em cima do endereço do local selecionado, conforme mostrado na Figura 4.

Figura 4 – Mapa do Google Maps com seleção de endereço



Fonte: Google Maps.

4. Na caixa de diálogos que surge, selecione “O que há aqui?”. Será mostrada outra caixa com algumas informações, inclusive a latitude e longitude do endereço selecionado, conforme mostra a Figura 5.

Figura 5–Mapa do Google Maps com seleção de endereço e visualização das coordenadas do endereço selecionado



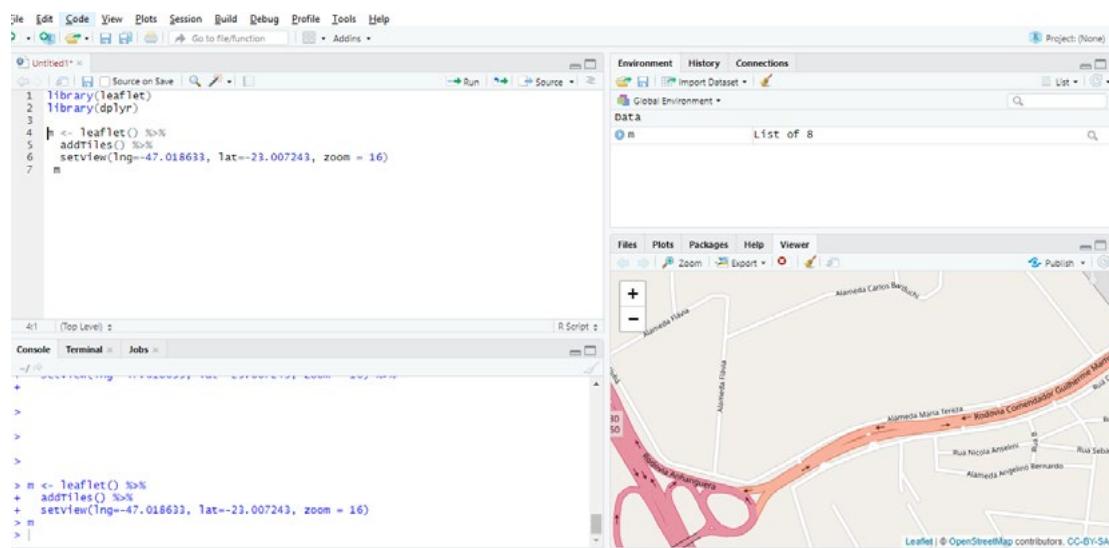
Fonte: Google Maps.

5. Anote a latitude e a longitude que aparece na Figura 5 mostrada.
6. Agora é possível começar a programar, conforme apresentado a seguir. Maiores detalhes sobre cada linha de comando podem ser encontrados em Tran (2019).

```
m-leaflet() %%
addTiles() %%
setView(lng=-47.018633, lat=-23.007243, zoom = 16)
m
```

7. O mapa é mostrado no RStudio no canto inferior direito, conforme mostra a Figura 6.

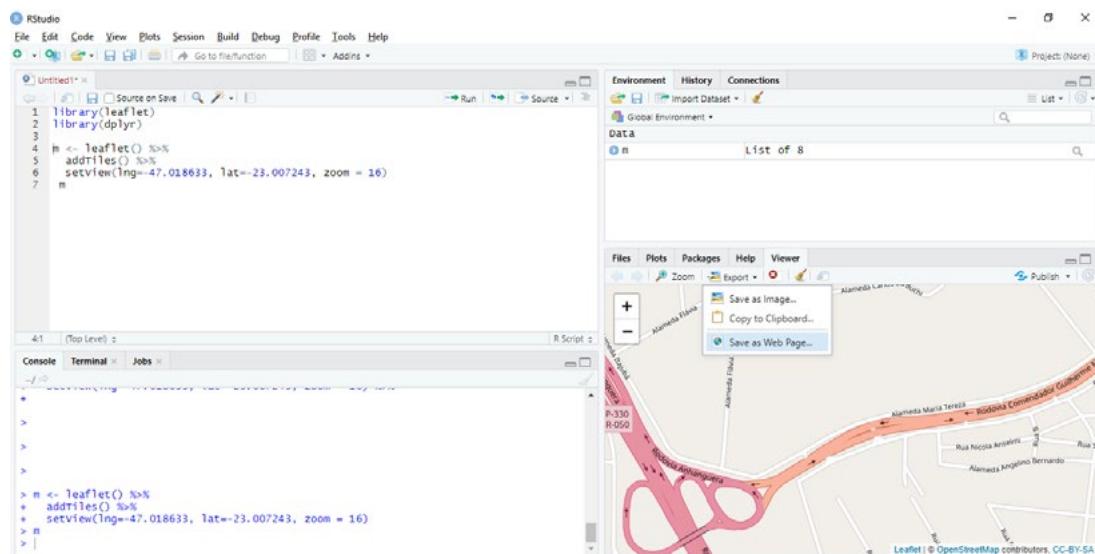
Figura 6 – Tela do RStudio com o mapa gerado



Fonte: elaborada pelo autor.

8. É possível exportar o mapa para a Web com a execução dos seguintes passos. Primeiro, clicar no botão “Export” e selecionar “Save as Web Page...”, como mostra a Figura 7.

Figura 7 – Tela do RStudio com preparação para exportação do mapa gerado.



Fonte: elaborada pelo autor.

9. Salvar com um nome de sua preferência, como, por exemplo, “corporativo-kroton.html”, com a extensão html e em uma pasta de seu computador de sua preferência.
10. Em seguida abrirá o mapa no seu navegador de uso.

É possível utilizar a biblioteca Leaflet fora do R também. No entanto, vai exigir do usuário um maior conhecimento de linguagem JavaScript, além de ter uma interface menos amigável que a interface do RStudio.

► 4. Datawrapper

Criada por um grupo de profissionais da área de tecnologia de *softwares*, a plataforma Datawrapper é de origem alemã. É uma ferramenta *online* criada para construir gráficos interativos. É muito simples e fácil de ser usada e produz desde gráficos e infográficos simples até os mais complexos possíveis.

A ferramenta tem um facilitador quanto à imputação dos dados que serão plotados, pois é possível inserir na própria plataforma *online* ou exportar os dados de uma planilha, como, por exemplo, a Microsoft Excel.

Uma das vantagens é que não exige conhecimento algum de linguagem de programação, além de fornecer direitos autorais completos para o usuário elaborador de alguma visualização.

A plataforma *Datawrapper* tem versões distintas, as quais estão relacionadas com o tipo de inscrição do usuário. Existe uma versão gratuita com menos facilidades que as versões pagas e, também, há assinaturas que permitem que mais de um usuário faça uso da plataforma.

Ao acessar o portal no endereço <https://www.datawrapper.de/>, o usuário terá acesso a uma série de informações e tutoriais que o ajudarão a desenvolver gráficos e mapas diversos de seu interesse.

► 5. Dygraphs

Biblioteca desenvolvida com uso de linguagem JavaScript e produz seus gráficos com o uso do canvas do HTML5. Assim como as bibliotecas JavaScript apresentadas neste texto, também possui código aberto (*open source*) e, na página da Dygraphs (<http://dygraphs.com/>), é afirmado que é considerada rápida e flexível.

Existe um pacote dygraphs em linguagem R que pode ser manipulado para produção de gráficos pelo RStudio. Sua limitação ou desvantagem é que produz unicamente gráficos de linhas.

Em Machado Neto (2013) é possível encontrar exemplos de aplicação com apresentação de *script* e de seu resultado gráfico. É possível, também,

encontrar em <https://rstudio.github.io/dygraphs>, a documentação do pacote dygraphs para R, assim como exemplos de aplicação.

► 6. Highcharts

Biblioteca de gráficos multiplataforma desenvolvida desde 2009, baseada em SVG (*Scalable Vector Graphics*), que, em português, é conhecida como gráficos vetoriais escalonáveis. Bastante utilizada para a elaboração de *dashboards* por apresentam um conjunto amplo de gráficos que podem ser gerados.

Considerada uma biblioteca responsiva pelos desenvolvedores (Shahid, 2014), por isso afirmam que funciona de forma eficiente tanto para a Web quanto para mobile. Se for utilizada para fins comerciais, exige um licenciamento, caso contrário, isso não se faz necessário.

Para ser utilizada é necessário realizar *download* no portal do seu desenvolvedor Highcharts e escolher a biblioteca que deseja trabalhar. Exige um certo conhecimento de linguagem de programação JavaScript.

► 7. Google charts

Biblioteca do Google desenvolvida em JavaScript adequada para a elaboração de gráficos para a Web e para dispositivos móveis. Segundo a descrição na página da ferramenta, afirma-se que a biblioteca é simples, de uso livre e fácil de ser utilizada (GOOGLE CHARTS), apesar de exigir conhecimento em linguagem JavaScript, HTML, CSS e algum editor de texto.

Assim como algumas bibliotecas aqui apresentadas, é possível executar e renderizar gráficos simples ou dinâmicos para Web ou dispositivos

móveis através de bibliotecas da linguagem R, como, por exemplo, o pacote Shiny, que se comunica facilmente com a linguagem JavaScript e produz visualizações diversas. Um exemplo de aplicação do uso do pacote Shiny com a biblioteca Google charts é apresentado por Cheng (2019), o qual mostra o resultado gráfico obtido e o *script* utilizado para a sua renderização.

ASSIMILE



HTML é uma linguagem de marcação, ou seja, é utilizada para a construção de páginas na Web, sua função é dizer como uma página Web deve ser estruturada, portanto, não é considerada uma linguagem de programação. Dentro da linguagem R, um formato HTML pode conter componentes interativos através de comandos *htmlwidgets*, que são comandos que produzem visualizações interativas em HTML (WICKHAM; GROLEMUND, 2017).

► 8. Polymaps

Biblioteca JavaScript de código aberto (*open source*), criada para renderização de mapas interativos e dinâmicos para visualização na Web. Na página da ferramenta (<http://polymaps.org/>) há uma descrição que afirma que o seu uso ajuda a apresentar visualizações com multizoom de conjuntos de dados em mapas de forma rápida por fazer uso de SVG, o que permite utilizar linguagem CSS para definir o *design* dos dados a serem apresentados.

Ainda na página da biblioteca, existem exemplos de utilização com os *scripts* disponibilizados e ampla documentação a seu respeito. No

mesmo local, também é possível fazer *download* da versão atualizada. A biblioteca Polymaps é um projeto sob responsabilidade da SimpleGeo e Stamen.

► 9. Weka

Diferente das demais ferramentas apresentadas neste texto, o Weka é um aplicativo apropriado para mineração de dados, tratamento de *big data* e aprendizagem de máquina. Trabalha especificamente com ferramentas de classificação e regressão, clusterização e identificação de associação entre dados.

O Weka foi criado na Universidade de Waikato, na Nova Zelândia, em 1997 e, desde lá, vem sendo aperfeiçoado e utilizado por uma ampla comunidade de profissionais de tecnologia da informação (TI). O nome Weka vem de Waikato Environment for Knowledge Analysis.

Na página do *software* (<https://www.cs.waikato.ac.nz/ml/weka/index.html>) há uma descrição sobre o Weka, que afirma ser uma coleção de algoritmos de aprendizagem de máquina (*machine learning*) para execução de mineração de dados.

O Weka tem em comum algumas ferramentas descritas neste texto e também é desenvolvido com linguagem Java e é de código aberto (*open source*). No entanto, tem a vantagem de utilizar uma GUI (*Graphical User Interface*), ou seja, possui uma interface gráfica própria, de certa forma, bastante amigável para o seu usuário, o que facilita bastante a implementação de atividades de mineração de dados. Apesar de ter GUI própria para a implementação de comandos, assim como nas demais bibliotecas requer o uso de um editor de textos, como, por exemplo, do Notepad++, para criação de *scripts* em linguagem Java. Portanto, exige um certo conhecimento desta linguagem de programação. A GUI do Weka é mostrada na Figura 8.

Figura 8 – Interface gráfica do Weka



Fonte: Elaborada pelo autor.

A página do Weka possui documentação e, também, disponibiliza diferentes versões para *download*. A GUI principal do Weka apresenta quatro modos de trabalho. Uma das mais utilizadas é a Weka Explorer, acessada através do botão explorer da GUI principal.

PARA SABER MAIS



No site do Weka é possível encontrar cursos gratuitos *online* com conteúdos diversos, como aprendizado de máquina e mineração de dados. Também é possível encontrar bastante material sobre Weka na internet, assim como vídeo aulas na plataforma Youtube.

Esta leitura fundamental apresentou um conjunto de ferramentas disponíveis para elaboração de visualização de dados. Cada uma com especificidades e disponibilidades variadas. Portanto, cabe ao interessado em produzir visualização, a ferramenta que achar mais útil para o desenvolvimento de seu trabalho, seja, corporativo ou acadêmico.



TEORIA EM PRÁTICA

Considere que você trabalha em uma empresa do ramo financeiro, ou seja, em uma consultoria financeira. O propósito de sua empresa é orientar potenciais investidores em investimentos diversos e prestar serviços de consultoria, com o intuito de fazer com seus clientes obtenham o melhor retorno possível.

Imagine que ocorrerá uma feira de negócios e você fica responsável por realizar uma apresentação dos serviços de sua empresa e também mostrar através de recursos computacionais os melhores investimentos dos últimos cinco anos.

Para elaborar a sua apresentação na feira, você pretende buscar uma ferramenta que te ajude a produzir gráficos atrativos, dinâmicos e fáceis de serem compreendidos. Afinal, o público potencial a ser expectador de sua apresentação poderá ser diverso.

Uma pergunta que vem em sua cabeça é “Qual ferramenta utilizar para produzir os gráficos? Dado que não tenho grande conhecimento de linguagem de programação e meu tempo de preparação é curto!”

Você, então, parte para a busca de algum recurso computacional que possa ajudar você a produzir uma boa apresentação e, assim, atrair mais clientes para o seu negócio.



VERIFICAÇÃO DE LEITURA

1. Muitas ferramentas de visualização de dados para web, conhecidas como bibliotecas, são construídas com base em uma linguagem de programação comum. De qual linguagem de programação estamos nos referindo?

Assinale a alternativa CORRETA.

- a. C++
- b. Java.
- c. JavaScript.
- d. Delphi.
- e. SAS.

2. A produção de visualização de dados para web está baseada no que se conhece por tríade de tecnologias. Qual dos elementos dessa tríade é responsável pela apresentação das páginas web?

Assinale a alternativa CORRETA.

- a. CSS.
- b. JavaScript.
- c. HTML.
- d. Web.
- e. HTML5.

3. A biblioteca Chart.js, desenvolvida por Nick Dowine, sob a linguagem JavaScript pode renderizar, ou seja, produzir quantos tipos de gráficos?

Assinale a alternativa CORRETA.

- a. Sete.
- b. Oito.
- c. Dez.
- d. Cinco.
- e. Nove.

Referências Bibliográficas

- CHENG, J. **Health expenditure vs. life expectancy**, 1995. Disponível em: <https://shiny.rstudio.com/gallery/google-charts.html>. Acesso em: 19 ago. 2019.
- CINTRA, J. **Gráficos comerciais na web com chart.js**. 2018. 21 slides. Disponível em: <http://josecintra.com/blog/wp-content/uploads/2018/11/chartjs.pdf>. Acesso em: 31 jul. 2019.
- FLANAGAN, D. **JavaScript**: o guia definitivo. 6. ed. Porto Alegre: Bookman, 2013. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788565837484/cfi/0!/4/2@100:0.00>. Acesso em: 31 jul. 2019.
- GOOGLE CHARTS. **Display live data on your site**. Disponível em: <https://developers.google.com/chart/?hl=pt-US>. Acesso em: 31 jul. 2019.
- MACHADO NETO, O. P. **Análise de bibliotecas para geração de gráficos na WEB**. 2013. 75 f. Trabalho de conclusão de curso (bacharelado em ciência da computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/86642/000910051.pdf;sequence=1>. Acesso em: 31 jul. 2019.
- SHAHID, B. **Highcharts essential**: create interactive data visualization charts with Highcharts JavaScript library. [N.p.]: Packt Publishing. 2014
- TRAN, A. B. **Interactive maps with leaflet**, 2019. Austin: University of Texas. Disponível em: https://journalismcourses.org/courses/RC0818/leaflet_maps.pdf. Acesso em: 31 jul. 2019.
- WICKHAM, H.; GROLEMUND, G. **R for data Science**: import, tidy, transform, visualize, and model data. Sebastopol: O'Reilly, 2017.

Gabarito

Questão 1 – Resposta: C.

Resolução: Muitas ferramentas de visualização de dados para web, conhecidas como bibliotecas, são elaboradas com base em uma linguagem de programação em comum, que é a JavaScript.

Feedback de reforço: Lembre-se que a linguagem JavaScript é bastante utilizada para produzir visualização de dados para web.

Questão 2 – Resposta: A.

Resolução: O componente do que se conhece como tríade de tecnologias responsável pela apresentação das páginas Web é o CSS.

Feedback de reforço: Lembre-se que a tríade de tecnologias é formada pela JavaScript, HTML e CSS.

Questão 3 – Resposta: B.

Resolução: A biblioteca Chart.js pode renderizar oito tipos de gráficos.

Feedback de reforço: Lembre-se das características da biblioteca Chart.js.



Bons estudos!