Visualização de Dados Aplicada

2024-09-23

Alunos: Bianca Lang, João Victor Pietchaki Gonçalves, Lenardo Eizo Sakai, Victor Pignatari

João Victor Pietchaki Gonçalves Lenardo Eizo Sakai Victor Pignatari Bianca Lang

Prof. Dr. Anderson Ara

Desafio #01 - CM303

DEST/UFPR

Objetivos

Este artigo visa dar um panorama sobre o que é a efetividade em visualização de dados bem como relatar 3 tipos de gráficos não-efetivos.

Visualização de Dados Aplicada

A visualização de dados não é um campo de aplicação recente. Suas origens datam de séculos atrás com a cartografia (século XVI), William Playfair (século XVIII) e posteriormente com Charles Minard (século XIX), por exemplo, e eram feitas manualmente em sua maioria.

Atualmente, grande parte das visualizações de dados se encontram em formas de dashboards interativos, infográficos e/ou até gráficos mais criativos. A flexibilidade trazida pelas tecnologias do século XXI (como o próprio Python, R e Power BI) permitiram um enorme avanço nessa área, garantindo ainda a abstração de ideias e padrões contidas em uma quantidade de dados cada vez maior.

Dada a grande quantidade e instantaneidade das informações, a construção de visualização de dados tem sido cada vez mais demandada. Entretanto, todas elas são de qualidade? Ou seja, todas são efetivas para com a principal função a qual deveria exercer?

O Que é um Gráfico Efetivo?

Por ter se tornado popular no ambiente de negócios e na comunicação jornalística, muitas e muitas visualizações têm sido criadas diariamente. Concomitantemente, nem todas comunicam ideias sobre o respectivo conjunto de dados de forma eficiente. Afinal, o que seria um gráfico efetivo?

Um gráfico efetivo, por força, é identificável. Não à toa, o célebre estatístico Edward Tufte é autor da seguinte frase: “Excelência gráfica é a que oferece ao espectador o maior número de ideias no menor tempo possível, com menos tinta no menor espaço”. Ademais, o mesmo admirável autor destaca 4 princípios para gráficos de excelência:

1. Acima de tudo, mostre os dados;
2. Maximize a taxa de dados e tinta;
3. Apague a tinta que não é de dados;

4. Apague a tinta de dados redundantes.

Junto a isso, Alberto Cairo, outro ilustre profissional de visualização de dados, nos indica o seguinte: "Os gráficos de informação devem ser esteticamente agradáveis, mas muitos designers pensam em estética antes de pensarem na estrutura em si, na história que o gráfico deve contar".

Ainda, e igualmente importante, a fidelidade aos dados é crucial para se construir gráficos efetivos! Aliás, como o próprio nome já diz, a visualização é dos dados, e não de ideias subjetivas de quem o desenvolve.

Sendo assim, tomaremos como base esses três grandes fundamentos (1) otimização de elementos gráficos para destacar informações, 2) **storytelling**, 3) fidelidade aos dados) e relataremos gráficos não-efetivos que encontramos em nossas pesquisas.

Exemplos de Gráficos Não-Efetivos

Mapas de Calor: Vendo Apenas em Absoluto

Figura 1 : Mapa de densidade populacional dos Estados Unidos. Fonte: Flowingdata.

Como editado em um artigo da Wikipédia, "um **mapa de calor** é uma técnica de visualização de dados que mostra a magnitude de um fenômeno por meio de cor em duas dimensões. A variação de cor pode ser por matiz ou intensidade, dando pistas visuais óbvias ao leitor sobre como o fenômeno está agrupado ou varia no espaço".

Seja o seguinte conjunto de dados um exemplo: números absolutos de roubos por estado nos Estados Unidos da América em 2024. Observando um gráfico que represente tal conjunto de dados, não é possível dizer que um estado é mais perigoso que outro justificando que o primeiro tem 2 roubos em um dado período de tempo enquanto o segundo apenas 1 registrado. E se o primeiro estado tiver 1000 vezes a população do segundo? Assim, observa-se que representar graficamente tais valores (absolutos) em um mapa de calor seria o mesmo que fornecer um mapa de densidade populacional, tornando impossível uma investigação comparativa entre regiões geográficas.

Para tornar *storytelling* possível, é muito mais útil pensar os dados em termos de porcentagens e taxas do que de valores totais ou absolutos.

Abaixo, uma ilustração cômica compactando bem a inefetividade de mapas de calor para valores absolutos:

Figura 2: História em quadrinhos sobre mapas de calor. Fonte: xkcd.

Na **figura 2**, o mapa do canto superior esquerdo representa os usuários de um site, o do canto superior direito os inscritos no *"Martha Stewart Living"*, e o último os consumidores de um certo tipo de pornografia. Entretanto, visualmente, os três mapas de calor são idênticos, como se fossem um mapa da densidade populacional dos EUA.

Como diz o texto da **figura 1**... **Vendo apenas em absolutos**: Isto é apenas população. Quando comparando entre lugares, categorias ou grupos, deve ser comparado de forma justa e considerar valores relativos

Gráfico de Área Empilhada Concêntrica: Confundindo Dimensão e Formato de Áreas

Figura 3: Ilustração para indicar diferenças de representação de área através de proporção e formato adequados. Fonte Flowingdata.

Na **figura 3**, vê-se que ambos os retângulos completam a mesma quantidade de área, mas têm formatos muito diferentes.

Quando falamos de representações gráficas que dispõem de áreas, é importante que elas sejam proporcionais aos valores que representam, bem como tenham o mesmo ou muito semelhante formato. Segue abaixo um gráfico de área empilhada concêntrica que "informa" a distribuição populacional do quanto é gasto em mercado por semana.

Figura 4: Gráfico de Área Empilhada Concêntrica do percentual de pessoas por faixa de gastos com mercado por semana. Fonte: Old Street Solutions.

Traduçao da **figura 4**:

Quanto você gasta em mercado por semana?

- 22% abaixo de $100
- 26% em torno de $100
- 39% de $100 a $200
- 10% de $200 a $300
- 3% mais que $300

Neste tipo de gráfico, as áreas maiores são sobrepostas pelas áreas menores. Entretanto, essa sobreposição torna desproporcionais as áreas e os formatos, causando muita confusão. Neste exemplo, fica nítido que aqueles que gastam de $100 a $200 representam a maioria da população, mas não é o mesmo para aqueles que gastam em torno de $100 e compõem o segundo maior grupo de pessoas. Ademais, a diferença entre as áreas amarela e azul é de apenas 4%, mas a azul aparenta ser muito maior. Não o bastante, esses 4% que de fato representam a cor amarela do gráfico aparentam ser muito maiores que o pequeno quadrado vermelho que representa 3% da população. Enfim, apenas um não para essa visualização!

Sobrepor áreas como esse gráfico torna desproporcional e gera deformação, impedindo uma visualização efetiva das informações. Uma escolha garantida seria o gráfico de barras, que permite exibir claramente as diferenças entre cada categoria. Aleternativamente, se o foco for a proporção em relação ao todo, também é possível utilizar o gráfico de setores, considerando que não são muitas observações, e possuem diferenças notáveis entre cada uma.

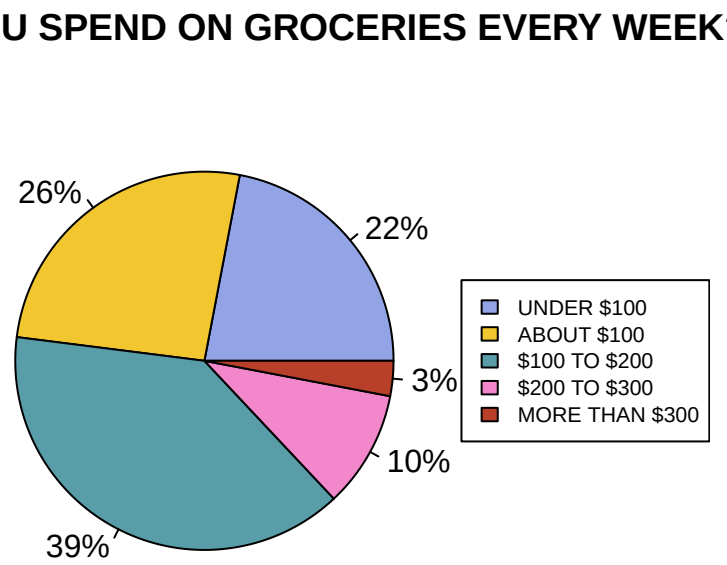

Figura 5: Gráfico de setores apresentando os mesmos valores do gráfico da figura 4. Fonte: De autoria própria.

Gráfico de Eixos Duplos: Criando Ideia de Causalidade a Partir de Correlação

Figura 6: Ilustração de Gráfico de Dois Eixos Verticais. Fonte: Flowingdata.

Um gráfico de dois eixos verticais usa duas escalas diferentes e pode ser um argumento de causalidade forçada. Ao usar eixos duplos, a magnitude pode encolher ou expandir para cada métrica. Isso é feito tipicamente para implicar correlação e causalidade. "Por causa disso, essa outra coisa aconteceu. Veja, está claro." Como não é senso comum que correlação não implica logicamente em causalidade e que a mera coincidência é tida por alguns como fonte para criação de teorias da conspiração, esse tipo de gráfico pode gerar confusão no sentido da integridade dos dados e da história a ser contada.

O exemplo abaixo foi resultado de um projeto de identificação de correlações espúrias conduzido por Tyler Vigen (ele descobriu mais de 4000 correlações com dados dos EUA. Curiosidade inútil: muitas coisas se correlacionavam com o consumo de queijo lol)

Figura 7: Taxa de divórcio a cada 1000 pessoas na cidade de Maine (EUA) se correlaciona com o consumo de libras *per capita* de margarina entre os anos 2000 e 2009. Fonte: Flowingdata.

Este foi um dos resultados que Tyler Vigen encontrou ao escrever um programa de computador para identificar automaticamente coisas que se correlacionassem. Como pode ser visto, com o passar dos anos, o número de divórcios a cada 1000 pessoas diminui de forma proporcional ao consumo de margarina nos EUA (com índice de correlação de *Pearson* a 0.992558), indicando que estariam na mesma tendência. Entretanto,

é verdade que quanto menos libras de margarina consumidas pela população dos EUA, menor a taxa de divórcio em uma cidade americana chamada Maine? Se ninguém mais nos EUA consumir margarina, então não existirá mais divórcios na respectiva cidade? Essas questões são possíveis de serem feitas a partir dessa visualização, entretanto, na realidade, sabemos que é uma mera coincidência!

De qualquer forma, o modo satírico e divertido de T. Vigen levantou alguns pontos importantes:

- Ser crítico com as estatísticas (e visualizações de dados em gráfico de dois eixos) que se vê;
- Saber que correlação não implica em causalidade;
- Exigir rigor científico para mostrar que há uma correlação forte e estatisticamente significativa.

Fontes

- YAU, Nathan. **How to spot visualization lies**. Disponível em: https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/. Acesso em: 22 set. 2024.
- YAU, Nathan. **Random things that correlate**. Disponível em: https://flowingdata.com/2014/05/12/random-things-that-correlate/. Acesso em: 22 set. 2024.
- FLETCHER, James. **Spurious correlations: Margarine linked to divorce?**. BBC News Magazine, 2014. Disponível em: https://www.bbc.com/news/magazine-27537142. Acesso em: 22 set. 2024.
- XKCD. **Correlation**. Disponível em: https://xkcd.com/1138/. Acesso em: 22 set. 2024.
- DUNFORD, Cristopher. **When Data Visualization Really Isn't Useful (and When It Is)**. Disponível em: https://www.oldstreetsolutions.com/good-and-bad-data-visualization. Acesso em: 22 set. 2024.
- WIKIPÉDIA. **Mapa de calor**. Disponível em: https://pt.wikipedia.org/wiki/Mapa_de_calor. Acesso em: 22 set. 2024.
- ARA, Anderson. (2024). Visualizaçao de Dados Aplicada Slides 02 [Slides]. Universidade Federal do Paraná.