

Visualização de Dados Aplicada

2024-09-23

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

Desafio #01 - CE303 - Prof. Dr. Anderson Ara - DEST/UFPR - 2024/2

Exemplos de Gráficos Não-Efetivos

Mapas de Calor: Vendo Apenas em Absoluto

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

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.

Aqui, uma ilustração cômica compactando bem a inefetividade de mapas de calor para valores absolutos.

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

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

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!

Sobrepôr 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. Alternativamente, 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.

HOW MUCH DO YOU SPEND ON GROCERIES EVERY WEEK?

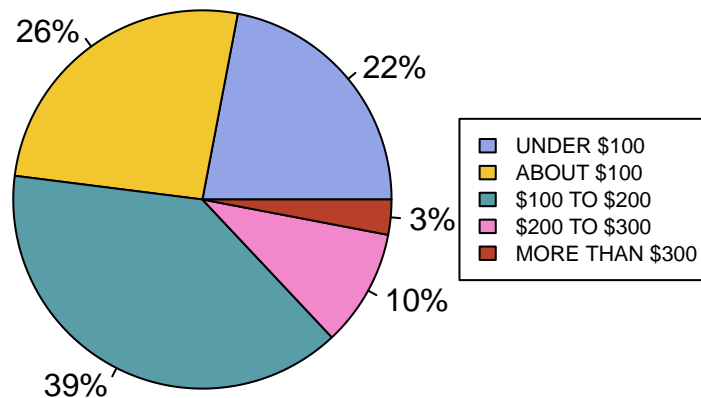


Figura 3: Gráfico de setores apresentando os mesmos valores do gráfico da figura 4. Fonte: De autoria própria.

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

Figura 4: 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. 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.

Figura 5: 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, indicando que estariam na mesma tendência. Entretanto, 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.