Pense comigo, se são tantos dados produzidos diariamente, por que não aprender com eles? É pensando nisso, que a Ciência de Dados busca gerar conhecimento a partir de dados. Mas calma, primeiro, é preciso compreender as informações que os conjuntos de dados representam, pois observando-os diretamente nem sempre é uma tarefa fácil, e se torna ainda mais difícil conforme a quantidade de dados aumenta.

Assim, para lidar com esses dados, é preciso analisá-los e interpretá-los. Por isso, nesta aula, você será apresentado à Visualização de Dados, que é a tarefa de representar dados, e é uma das principais habilidades de análise, pois amplia a nossa capacidade de compreender e explorar grandes conjuntos de dados. Dessa forma, você conhecerá alguns dos princípios básicos para transformar dados em imagens que sejam verdadeiras, belas, funcionais e reveladoras. Além, é claro, de entender a importância de tudo isso.

**Objetivos**

* Entender a área de visualização de dados e a sua relevância no contexto da ciência de dados;
* Identificar as boas qualidades de uma visualização;
* Conhecer os princípios para mapear dados em elementos gráficos.

**Tópico 1 – Introdução à Visualização de Dados**

**OBJETIVOS**

* Compreender a importância de visualização de dados;
* Entender a definição de visualização de dados;
* Analisar exemplos de uma visualização.

Após aquela breve introdução, me diga, você consegue observar uma tabela cheia de números e rapidamente ter boa noção sobre as informações que estão contidas ali? Para tabelas com poucas linhas, isso pode ser possível, não é mesmo? Porém, conforme você vai crescendo, o tamanho dos dados, esse tipo de análise, se torna inviável até para profissionais experientes. Por isso, não é preciso se apavorar. Segundo pesquisadores, isso é apenas uma característica do cérebro humano, viu? Ocorre que a maioria dos cérebros humanos simplesmente não são bons em processar longas e monótonas sequências de números ou textos. Contudo, eles são bem treinados para processar e reconhecer padrões visuais. Neste tópico, você aprenderá o que é visualização de dados e por que ela é uma ferramenta poderosa para explorar e compreender conjuntos de dados. Ah, você também irá aprender a identificar o que faz com que uma visualização seja boa.

Vamos lá?!

Para começar, vamos fazer um experimento. A seguir, há quatro tabelas – 1, 2, 3 e 4 – com dados fictícios sobre sites de compras. Cada tabela representa um site diferente e é composta por três colunas: **Usuário**, que diz respeito a cada usuário que acessa o site; **x**, que representa a quantidade de vezes que o usuário fez login no site; e **y**, que indica a quantidade total de horas que o usuário passou logado.

| **Tabela 1** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Usuário** | **x** | | | | **y** |
| A1 | 10 | | | | 8 |
| B1 | 8 | | | | 7 |
| C1 | 13 | | | | 8 |
| D1 | 9 | | | | 9 |
| E1 | 11 | | | | 8 |
| F1 | 14 | | | | 10 |
| G1 | 6 | | | | 7 |
| H1 | 4 | | | | 4,26 |
| I1 | 12 | | | | 10,84 |
| J1 | 7 | | | | 5 |
| K1 | 5 | | | | 6 |
| **Tabela 2** | | | | | |
| **Usuário** | | | **x** | | **y** |
| A2 | | | 10 | | 9,14 |
| B2 | | | 8 | | 8,14 |
| C2 | | | 13 | | 8,74 |
| D2 | | | 9 | | 8,77 |
| E2 | | | 11 | | 9,26 |
| F2 | | | 14 | | 8,1 |
| G2 | | | 6 | | 6,13 |
| H2 | | | 4 | | 3,1 |
| I2 | | | 12 | | 9,13 |
| J2 | | | 7 | | 7,26 |
| K2 | | | 5 | | 4,74 |
| **Tabela 3** | | | | | |
| **Usuário** | | **x** | | **y** | |
| A3 | | 10 | | 7,46 | |
| B3 | | 8 | | 6,77 | |
| C3 | | 13 | | 12,74 | |
| D3 | | 9 | | 7,11 | |
| E3 | | 11 | | 7,81 | |
| F3 | | 14 | | 8,84 | |
| G3 | | 6 | | 6,08 | |
| H3 | | 4 | | 5,39 | |
| I3 | | 12 | | 8,15 | |
| J3 | | 7 | | 6,42 | |
| K3 | | 5 | | 5,73 | |

| **Tabela 4** | | |
| --- | --- | --- |
| **Usuário** | **x** | **y** |
| A4 | 8 | 6,58 |
| B4 | 8 | 5,76 |
| C4 | 8 | 7,71 |
| D4 | 8 | 8,84 |
| E4 | 8 | 8,47 |
| F4 | 8 | 7,04 |
| G4 | 8 | 5,25 |
| H4 | 19 | 12,5 |
| I4 | 8 | 5,56 |
| J4 | 8 | 7,91 |
| K4 | 8 | 6,89 |

Observando as tabelas, você consegue dizer algo sobre os sites 1, 2, 3 e 4? Por exemplo, existe alguma relação entre a quantidade de vezes que os usuários fizeram login e o tempo que eles gastam nos sites? Ou, essa relação é a mesma para todos os sites? Estas são algumas perguntas que poderiam surgir ao se trabalhar com estes dados. Portanto, note que apenas conferir os números brutos pode ser muito cansativo e os padrões podem não ser evidentes. Por isso, as tabelas são úteis para consultar valores específicos, mas não para extrair informações gerais sobre os dados.

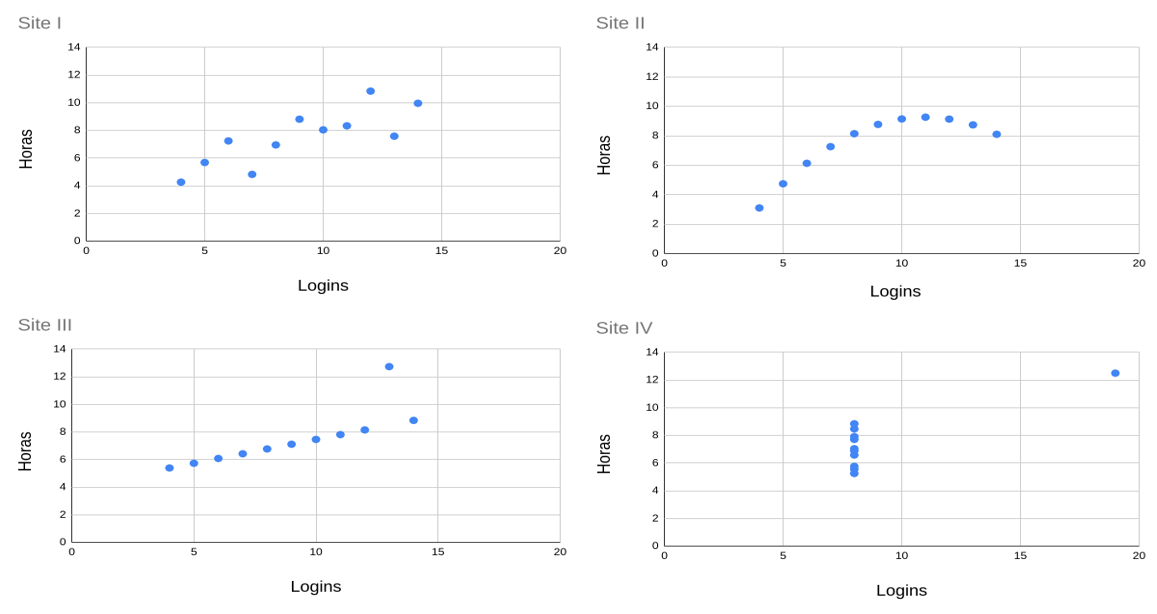
Para facilitar o entendimento dos dados, você pode calcular algumas medidas estatísticas, como média e variância. As tabelas a seguir são referentes às tabelas anteriores, mas apresentam apenas com o cálculo da média e do desvio padrão – que indica o quanto os dados de um conjunto de dados são semelhantes – de cada variável em cada site. Confira:

| **Tabela 1** | | |
| --- | --- | --- |
|  | **x** | **y** |
| **Média** | 9,00 | 7,50 |
| **Desvio padrão** | 3,32 | 2,03 |
| **Tabela 2** | | |
|  | **x** | **y** |
| **Média** | 9,00 | 7,50 |
| **Desvio padrão** | 3,32 | 2,03 |
| **Tabela 3** | | |
|  | **x** | **y** |
| **Média** | 9,00 | 7,50 |
| **Desvio padrão** | 3,32 | 2,03 |

| **Tabela 4** | | |
| --- | --- | --- |
|  | **x** | **y** |
| **Média** | 9,00 | 7,50 |
| **Desvio padrão** | 3,32 | 2,03 |

Você percebeu algo estranho nesses dados? Em cada site, a média e o desvio padrão da quantidade de visitas são idênticas, assim como o tempo gasto em cada um deles. Nessa perspectiva, podemos concluir que o comportamento dos usuários foi o mesmo em cada site? Talvez você concorde que essa pergunta é mais facilmente respondida quando os dados estão disponibilizados em um gráfico, como pode ser conferido na próxima página. Vamos lá?

Dados são geralmente representados em gráficos com um sistema de coordenadas que indica a posição no gráfico de cada determinado valor nos dados. Deu para entender? Vamos analisar o exemplo a seguir! Nele, existem duas variáveis, sendo cada uma delas representada em um eixo de um gráfico. Os usuários são representados por pontos organizados no espaço. Cada vez que o ponto estiver localizado mais para a direita (eixo horizontal x), significa que mais vezes o usuário logou no site; e quanto mais acima (eixo vertical, y), mais tempo o usuário passou logado. Portanto, esse tipo de visualização é chamado de **gráfico de dispersão**, ou scatterplot, tudo bem? Observe, na figura a seguir, a visualização de dados dos quatro sites apresentados em tabelas anteriormente:

Figura 1 – Gráfico de visualização de sites

Confira que, apesar de terem a mesma média e o mesmo desvio padrão, o comportamento dos usuários é completamente diferente em cada site. Esses diferentes padrões são percebidos mais rapidamente através de uma representação visual. Você já teve contato com esse tipo de gráfico antes? Esse conjunto de dados, famoso na estatística, chama-se Quarteto de Anscombe, e ilustra a necessidade de sempre organizar os dados para melhor entendê-los. Agora você percebe como isso é útil? Após entender esse experimento, vamos aprender um pouco mais sobre a Visualização de Dados!

Formalmente, nós podemos definir Visualização de Dados como a área do conhecimento dedicada à geração de imagens que auxiliam na compreensão de dados. Assim, o objetivo da Visualização é **ampliar** a capacidade humana de entender dados e não substituir as pessoas com processos automatizados, como muitas pessoas acreditam. Ela é especialmente útil quando não se sabe, de antemão, que tipo de pergunta se deseja responder sobre os dados. Boas visualizações dão uma noção geral e imediata sobre os dados, tornando explícitos padrões, tendências, exceções e comparação entre valores. Além disso, são ótimas ferramentas para comunicar informações sobre dados para o público, pois são mais facilmente interpretadas, entendeu? Você, provavelmente, já deve ter tido contato com vários tipos de visualizações. Alguns professores, por exemplo, usam gráficos em suas aulas para mostrar aos alunos o crescimento da população em determinada área; o desempenho da economia de um país; ou mesmo, o aumento de mortes que ocorreram em uma pandemia durante determinado tempo.

Você consegue lembrar de mais algum exemplo? Se sim, você, certamente, vai concordar que esses gráficos ajudaram a compreender melhor o assunto e a interpretar melhor os dados, não foi? Então, é possível dizer que isso ocorre porque, como já foi comentado, algumas pesquisas afirmam que o sistema visual humano é extremamente rápido e eficiente para identificar padrões, ou seja, o sistema visual percebe padrões de forma quase imediata e sem precisar de muito raciocínio.

Quer saber mais como isso funciona? Vamos fazer mais um teste a seguir!

A seguir, há uma sequência de números consideravelmente grande e, nela, o número 7 aparece algumas vezes. Só em conferir essa sequência, você consegue contar rapidamente quantas vezes o número 7 aparece? Verifique.

09181973892184998590887696980913895281798490812905891389859896842985283483912785987496382529524098502395809

Bastante difícil, não é mesmo? No entanto, se o número 7 estiver destacado com cor diferente na sequência? Será que muda alguma coisa? Confira.

091819**7**389218499859088**7**696980913895281**7**98490812905891389859896842985283483912**7**8598**7**496382529524098502395809

E aí? Com certeza, desse modo, esse desafio torna bem mais fácil a identificação do número 7, concorda? Perceba o quanto é importante organizar os dados da melhor forma para que eles possam ser interpretados também de maneira adequada. A seguir, continue acompanhando explicações e exemplos que indicam a melhor forma de se criar visualizações. Vamos lá?

Visualização de dados é uma área interdisciplinar, ou seja, pode-se dizer que é tanto arte quanto ciência, pois é preciso ter conhecimento tanto sobre dados, estatística e computação, como sobre design e estética, para criar visualizações que sejam tanto informativas como atrativas. Assim, boas visualizações devem ter as seguintes qualidades:

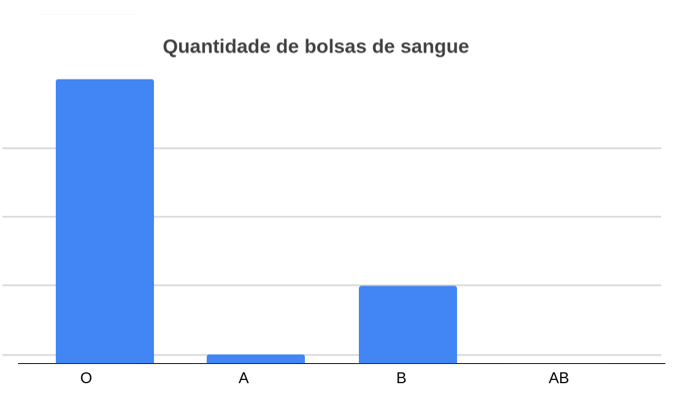
* **Verdadeira**: representar dados objetivamente e não induzir a conclusões falsas, pois uma visualização que não é verdadeira possui informações incorretas;
* **Funcional**: deve ser clara e de fácil compreensão. Uma visualização que não é funcional pode ser confusa ou possuir informação em excesso;
* **Bela**: ser esteticamente agradável. Uma visualização que não é bela pode ser clara e informativa, mas não atrai o observador. Pode não parecer uma qualidade importante, mas as visualizações que não são belas passam menos credibilidade e não focam a atenção, prejudicando o seu objetivo final.
* **Reveladora**: revelar informações sobre os dados. Uma visualização que não é reveladora, apenas replica os dados de uma tabela e não leva a descoberta alguma de conhecimento.

Até aqui você está conseguindo compreender? Pois bem, note que essas características estão fortemente inter-relacionadas. Por exemplo, uma imagem que não é funcional provavelmente não será bela, e vice-versa. A partir da próxima página, analise alguns exemplos que mostram como uma visualização deve ser verdadeira, funcional, bela e reveladora.

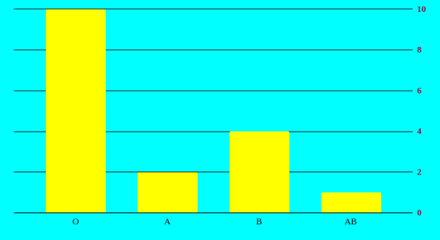
Para começar, observe a tabela a seguir com dados fictícios sobre a quantidade de bolsas de sangue coletadas em uma unidade de doação de sangue:

| **Tabela 5** | | |
| --- | --- | --- |
| **Tipo sanguíneo** | **Quantidade** |  |
| O | 10 |  |
| A | 2 |  |
| B | 4 |  |
| AB | 1 |  |

Observe que a tabela possui duas variáveis: uma categórica e uma quantitativa. Embora, nesta tabela, você possa observar todos os dados de uma vez, pode-se dizer que ela não é reveladora, pois não evidencia de forma rápida a relação entre as quantidades de bolsas de sangue. Por isso, uma forma bastante utilizada de visualizar quantidades de diferentes classes é através de gráficos de barras, nos quais a altura de cada barra representa a quantidade de cada classe. Confira, a seguir, algumas formas de construir gráficos de barras para representar a Tabela 5 e analise cada um de acordo com as qualidades apresentadas anteriormente.

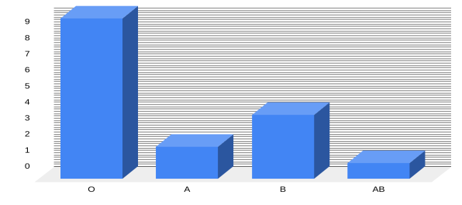
Figura 2 – Quantidade de bolsas de sangue – Visualização 1

Perceba que, na Visualização 1, as proporções entre as quantidades de bolsas de sangue estão alteradas e não há informações sobre as quantidades reais. Pela imagem, parece que há zero bolsas AB. Pode-se dizer que não é uma visualização verdadeira. A partir disso, saiba que é preciso estar atento para identificar visualizações que não são verdadeiras, pois elas são muito persuasivas e podem ser armas poderosas para manipular o público. No caso dessa imagem, o problema foi que o eixo vertical, que corresponde à quantidade das bolsas de sangue, não começa no valor zero. Note que a primeira linha guia que aparece seria a correspondente ao valor 2. Logo, a quantidade de bolsas AB não irá aparecer, pois equivale a 1.

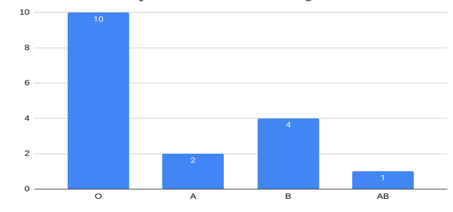
Figura 3 – Quantidade de bolsas de sangue – Visualização 2

A Visualização 2 é similar à anterior, possuindo as mesmas informações, mas possui problemas estéticos devido ao uso desnecessário de cores berrantes. Embora seja informativo e relativamente claro, não é atrativo para o leitor. Pode-se dizer que não é **belo**.

Vamos avaliar mais alguns gráficos?

Figura 4 – Quantidade de bolsas de sangue – Visualização 3

A Visualização 3 também contém as mesmas informações das anteriores, mas o uso excessivo de linhas guias e de representação em 3D torna o gráfico poluído, dificultando o entendimento. Pode-se dizer que não é uma visualização **funcional**.

Figura 5 – Quantidade de bolsas de sangue – Visualização 4

Agora, analise o gráfico de barras da figura 5. Ele é uma visualização mais adequada da Tabela 5. Ele é claro, objetivo e permite compreender de forma intuitiva a diferença entre as quantidades de bolsas de sangue, que é o seu propósito.

Interessante, não é mesmo? Você percebe o quanto a organização desses dados implica na forma como eles serão interpretados? Pois bem! Você ainda tem muito o que aprender sobre isso. A seguir, acompanhe as considerações gerais sobre o que foi apresentado até aqui.

É importante ter em mente que estas qualidades são guias gerais pela qual devemos nos orientar durante a criação de uma visualização. Não há regras rígidas para se obter visualizações com todas as boas qualidades, e algumas vezes é preciso sacrificar uma qualidade para favorecer outra, a depender do objetivo principal. Note ainda que as qualidades em si podem ser bastante subjetivas, e variar de acordo com o público para o qual a visualização é direcionada. Uma visualização que é funcional para cientistas pode não ser para o público leigo. A busca desse equilíbrio é o que torna Visualização de Dados parecida com uma forma de arte. Com estudo e prática, você irá desenvolver naturalmente a intuição para criar visualizações verdadeiras, funcionais, belas e esclarecedoras para diferentes contextos e públicos.

No próximo tópico você irá conhecer alguns dos princípios básicos de como mapear dados para imagens, buscando manter as boas qualidades de uma visualização. Ah, você também irá conhecer alguns tipos de gráficos mais utilizados e a quais tipos de dados eles melhor se adequam. Preparado(a)? Vamos lá!