

Definindo Visualização

Assim é como eu vou organizar essa aula. Primeiro de tudo, eu vou definir o que quero dizer com visualização. Então, vamos explorar os diferentes componentes de quaisquer visualizações de dados. Em nosso terceiro passo, vamos falar um pouco sobre como a visualização pode mentir. E isso está ligado ao meu mais recente livro "Como gráficos Mentem", ou melhor dizendo, como por vezes, mentimos para nós mesmos com os gráficos que vemos todos os dias. E então, finalmente, nós estamos aprender certos princípios muito importantes de design que se aplicam a visualização quando o propósito da visualização é comunicar ideias.

Então, vamos começar por definir o que entendemos por visualização de dados. Visualização de dados, como eu disse antes, basicamente consiste mapear dados em objetos. E quando esses objetos, quando os dados são mapeados para esses objetos, determinadas propriedades desses objetos vão variar de acordo com os dados que você está tentando representar. Nós vamos chegar a isso em apenas um minuto.

Mas o propósito, a razão pela qual fazemos isso, a razão porque visualizamos dados é porque queremos que as pessoas sejam capazes de extrair significado daqueles dados. A visualização de dados é, acima de tudo, uma ferramenta que podemos usar para extrair insights a partir dos números que vemos todos os dias, ou que vemos em um planilha ou em qualquer lugar. Certo? A maneira que eu costumo explicar isso é que uma visualização é como um par de óculos. Então, se eu tirar meus óculos, eu não sou capaz de ver as coisas muito claramente. Mas se eu colocar meus óculos, esta ferramenta que eu coloco na frente dos meus olhos me permite ver melhor.

A visualização de dados pode ser conceituada de uma forma muito semelhante. É uma ferramenta que nos permite ver além da complexidade dos dados. E quais são as finalidades da visualização de dados? Bem, a visualização de dados pode ser usada para explorar os dados, e você já aprendeu um pouco sobre isso em módulos anteriores deste curso. Mas também pode ser usada para comunicar ideias com base nesses dados, que é o que estamos focando no módulo atual.

Assim, a visualização pode ser utilizado para a exploração e descoberta, ou ele pode ser usada para a comunicação. Embora em ambos os casos, a ideia é, basicamente, a mesma. A ideia é que nós, seres humanos, temos muita dificuldade em identificar padrões e tendências de grandes quantidades de dados. Por exemplo, os dados que você tem na tela no momento é um grande conjunto de dados de temperaturas globais do ano 1000 até o ano 2000, medido em graus Celsius em comparação com uma média. A média do século 20 entre 1961 e 1980. Se não estou enganado. Certo? Então essa é a razão pela qual você verá valor negativo e os valores positivos neste conjunto de dados.

Se eu lhe mostrar esses dados, disse, se eu lhe der, por exemplo, a planilha Excel na qual os dados estão colocados e fizer perguntas com base nestes dados, será muito difícil para você responder essas perguntas. Por exemplo, vamos supor que eu mostro-lhe este dado, e pergunto, "A temperatura média global em 2000, no ano de 2000, foi superior ou inferior à do ano 1000?".

Bem, se você quiser responder a essa pergunta com base na conjunto de dados, você vai precisar para abrir a planilha Excel. Ir para o topo da planilha. Dê uma olhada na temperatura do ano 1000. Em seguida, role para baixo, todo o caminho, na planilha. Dê uma olhada no último número no conjunto de dados e em seguida, compare mentalmente - no seu cérebro - um número com o outro número e veja qual deles é mais elevado do que o outro.

Alerta de spoiler: A temperatura média em 2000 foi mais elevada do que a temperatura média de 1000. Certo? Mas a pergunta é difícil de responder. Estou forçando-o a abrir o conjunto de dados e, em seguida, vá de cima a baixo. Certo? É difícil.

Agora deixe-me perguntar uma questão ainda mais difícil. Vamos supor novamente que eu te mostre este conjunto de dados e pergunto, "Há qualquer ponto nos últimos 1000 anos em que a temperatura média foi maior do que era no ano de 2000, que é o último ano neste conjunto de dados particular?" Essa é uma pergunta muito mais difícil de responder.

Estou obrigando-o a abrir novamente o conjunto de dados, rolar de cima a baixo, ler cada número único, e ver se algum deles é maior do que o último número do conjunto de dados. Isso é muito difícil de fazer. Mas e se, em vez de lhe mostrar os dados, eu transformar, eu mapear esses dados em certos objetos. Certo? E eu transformo esses objetos na proporção dos dados que eu tenho no meu conjunto de dados. Em outras palavras, se eu transformar esses dados em um gráfico de linhas. Certo? Os dados que eu estava mostrando são os dados por trás de uma das mais famosas visualizações de dados concebidas durante o século 20. É comumente chamado de gráfico de taco de hóquei.

E a história que o gráfico de taco de hóquei nos diz é que no passado, entre o ano 1000 até o ano de 1900 mais ou menos, as temperaturas globais variaram. Essa seria a linha preta que você vê no meio. As temperaturas globais variaram, mas elas variam dentro de um determinado intervalo. E é só depois do ano de 1900 que as temperaturas globais começaram a subir muito, muito rapidamente. Certo? Essa é a história que o gráfico transmite. É o padrão que nós não vimos antes quando estávamos olhando para os números. Mas esse padrão torna-se evidente, torna-se visível, torna-se inevitável uma vez que mapeamos esses números em objetos, e transformamos todos esses números em uma gráfico específico.

Este gráfico, aliás, contém toneladas de camadas de informações. Primeiro de tudo, você vai ver que há uma linha azul. A linha azul são as nossas estimativas de temperaturas, certo, para todos os anos no passado. E eu digo estimativas porque, obviamente, não temos registros de temperaturas a partir do ano 1000 ou a partir do ano 1500. O que cientistas climáticos fazem para estimar a temperatura era no passado é para dar uma olhada em variáveis.

Variáveis que podemos explorar e que podem nos dizer algo sobre a variável que queremos estudar. Certo? Se não temos acesso direto à temperatura de um determinado ano, o que poderíamos fazer, por exemplo, é dar uma olhada em padrões de crescimento de anéis de árvores. Se você cortar uma árvore e você der uma olhada nos os anéis dentro, com base na largura e as distâncias entre esses anéis, você pode classificar de estimar o que temperatura que determinada árvore estava experimentando em cada ponto no tempo. Então, se você tem uma árvore muito antiga, você meio que pode ter um mapa das temperaturas globais da área onde que árvore cresceu. Ou qual era a temperatura nesse lugar específico.

Em qualquer caso, a linha azul é a estimativa baseada nestas variáveis. E depois a linha vermelha. A linha vermelha são os registros reais de temperaturas, começando, eu acredito, no século 18. Em meados do século 18, do século 19, temos registros reais de temperaturas, e essa é a linha vermelha. A linha preta que você vê no meio, que é chamada suave. Certo? É uma espécie de direção média dos dados. É uma linha que foi adicionada para tornar mais evidente qual a direção dos dados, se sobe ou desce. Porque, como você pode ver, as linhas azuis e vermelhas são muito onduladas. Certo? Eles vão para cima e para baixo assim. Parecem quase aleatórias.

E, em seguida, a área cinza por trás das linhas, que é o nível de incerteza dos dados. Obviamente, já que você está usando variáveis para estimar algo que você deseja analisar. Se você usar uma variável estimada, haverá um monte de incerteza em torno disso. Mas uma vez que você tem, que é a razão pela qual a área cinzenta é tão grande, tão grande no passado. Mas essa área cinzenta, esse nível de incerteza, torna-se mais estreito e mais estreito e mais estreito do na medida em que chegamos ao presente, porque uma vez que você tem registros reais de temperaturas, esse nível de incerteza torna-se muito, muito pequeno.

Este é o poder de visualização de dados. O poder de deixá-lo ver padrões, tendências, histórias que podem passar despercebidos se você só pode olha para uma linha de números ou para os números reais no seu conjunto de dados. Depois de transformar aqueles em algum tipo de gráfico, você será capaz de ver essas histórias imediatamente.