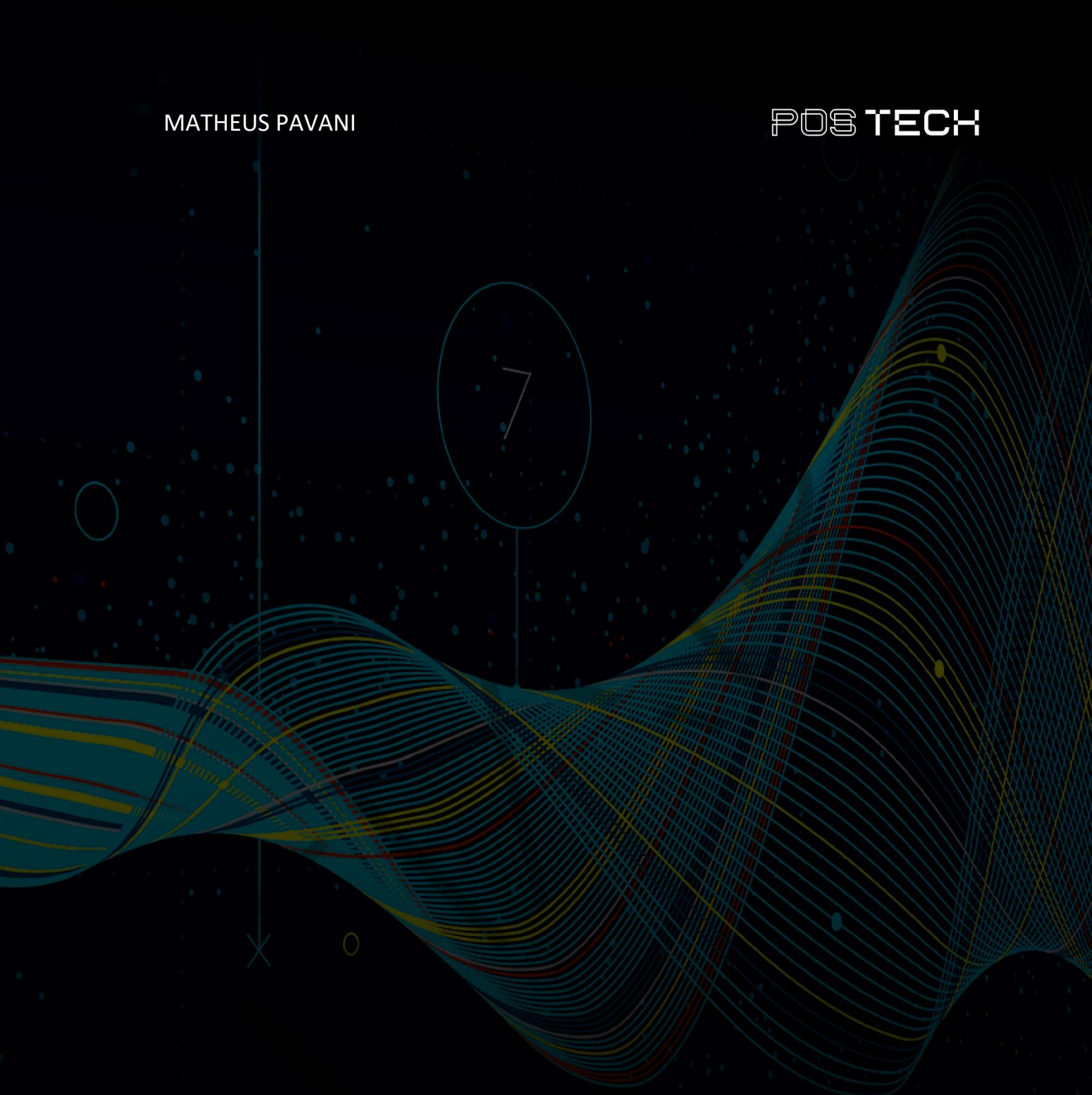
MATHEUS PAVANI

PÓS TECH



DATA ANALYTICS

ANÁLISE EXPLORATÓRIA DE DADOS

AULA 02

SUMÁRIO

O QUE VEM POR AÍ?

Nesta aula aprenderemos a como iniciar a nossa trajetória na visualização dos dados do Dataframe utilizando Matplotlib. Para cada situação, é importante que saibamos como lidar com os dados que temos em mãos. Não existe uma receita exata do que fazer, mas é preciso conhecer as técnicas principais.

Para isso, você conta com os materiais disponíveis e o canal do Discord para compartilhar ideias e dúvidas. Estamos aqui para te ajudar da melhor maneira possível!

As bases de dados podem ser baixadas como zip ou consultadas diretamente no GitHub.

IMPORTANTE: Não esqueça que temos um desafio para essa disciplina, que faz parte da sua jornada de aprendizado. Você pode acessá-lo na área de atividades da plataforma!

HANDS ON

Agora chegou o momento de ver, na prática, como começar a importação dos dados e trabalhar com eles via programação. O ambiente utilizado é o Google Colab, e as bases de dados que foram disponibilizadas no início deste material. A ideia aqui é não se limitar apenas ao código explícito no hands on, então o ideal é procurar a documentação das bibliotecas, explorar novas funcionalidades e muito mais.

O link para os notebooks utilizados estão em nosso GitHub para que você possa consultar e seguir com os seus estudos.

A ideia aqui é fortalecer seus conhecimentos técnicos, mas também os conteúdos mais analíticos descritos no texto.

SAIBA MAIS

Recapitulando...

A biblioteca Pandas é uma das mais completas quando o assunto é analisar dados, porque ela nos permite ler e gravar arquivos em diversas extensões (.csv, .xlsx, .parquet, .txt, .sas, .pkl, .html, .hdf , etc.), além de ler querys e tabelas em bancos de dados (desde que você conecte o Python no banco que desejar).

Já a biblioteca Matplotlib <https://matplotlib.org/> é a biblioteca mais usada para visualização de dados em forma de gráfico, além de ser base para a criação de outras bibliotecas (Seaborn <https://seaborn.pydata.org/> e Plotly <https://plotly.com/>). Aqui temos várias possibilidades no que diz respeito a criar e manipular gráficos, desde o tipo, até ajustes visuais como plano de fundo, eixos, tickets, títulos, legendas, subplots e a lista continua...

É importante que você estude os materiais disponíveis e assista as aulas, para aprender como podemos tratar esse conteúdo. Vamos começar?

DataViz

Cada vez mais o termo “DataViz” tem ganho um espacinho especial em nossos corações, pelo motivo que, em um mundo orientado a dados, é importante uma boa visualização para uma execução de insights de qualidade.

Este termo é uma abreviação de “Data Visualization”, ou seja, a forma pela qual representamos nosso conjunto de dados, sejam eles estruturados ou não.

Além do que veremos em aula, é interessante aprofundarmos o tema, para que não tenhamos o risco de nos perder no processo.

A visualização de dados é uma técnica que permite representar informações numéricas e estatísticas de forma gráfica, tornando-as mais fáceis de entender e interpretar. Essa técnica é amplamente utilizada em muitas áreas, incluindo negócios, ciências, saúde, tecnologia e governo, para ajudar a tomar decisões informadas e a compreender informações complexas.

Dessa forma, isso é importante porque nosso cérebro é muito mais eficiente em processar informações visuais do que dados brutos. Isso significa que é muito mais fácil identificar tendências, padrões e relações entre dados quando eles são representados de forma visual. Além disso, a visualização de dados permite comparar rapidamente vários conjuntos de dados e detectar anomalias ou outliers.

Vivemos em uma época de informação visual, e o conteúdo visual desempenha um papel importante em todos os momentos de nossas vidas. Uma pesquisa conduzida pela SHIFT Disruptive Learning, mostrou que geralmente processamos imagens 60.000 vezes mais rápido do que uma tabela ou texto e que nossos cérebros fazem um trabalho melhor de lembrá-los no futuro. O estudo constatou que, após três dias, os estudos analisados retinham entre 10% e 20% das informações escritas ou faladas, em comparação com 65% das informações visuais.

O cérebro humano pode perceber imagens em apenas 13 milissegundos e armazenar informações, desde que associadas ao conceito. Nossos olhos podem capturar 36.000 mensagens visuais por hora. 40% das fibras nervosas estão conectadas à retina.

Tudo isso mostra que as pessoas processam melhor as informações visuais, que estão embutidas em nossa memória de longo prazo. Como resultado, em relatórios e declarações, a representação visual usando imagens é uma forma mais eficaz de comunicar informações do que texto ou tabela; e ocupa pouco espaço. Isso significa que a visibilidade dos dados é mais atraente, fácil de interagir e fácil de lembrar.

Há muitas ferramentas e técnicas disponíveis para ajudar na visualização de dados, incluindo gráficos de barras, gráficos de linhas, gráficos de dispersão, mapas e infográficos. Cada tipo de visualização é adequado para diferentes tipos de dados e objetivos, e é importante escolher a visualização adequada para garantir que a informação seja transmitida de forma clara e precisa.

Além disso, a visualização de dados também é útil para a comunicação de informações. Por meio de gráficos e infográficos atrativos, é possível transmitir informações de maneira clara e memorável a públicos de diferentes contextos e níveis de conhecimento técnico. Isso pode ser especialmente útil em apresentações

comerciais, onde é necessário convencer outras pessoas a tomar uma decisão ou adotar uma perspectiva específica.

À medida que a “era do Big Data” entra em ação, a visualização é uma ferramenta cada vez mais importante para entender os trilhões de linhas de dados gerados todos os dias. A visualização de dados nos ajuda a contar histórias, organizando os dados de uma forma mais fácil de entender, destacando as tendências e os valores discrepantes. Uma boa visualização conta uma história, removendo o ruído dos dados e destacando informações úteis.

No entanto, não é tão fácil quanto apenas enfeitar um gráfico para torná-lo melhor, alterando a parte de “informações” de um infográfico. A visualização efetiva de dados é um ato de equilíbrio delicado entre forma e função. O gráfico mais simples pode ser muito complicado para chamar a atenção ou revelar um ponto importante; a visualização mais impressionante pode falhar totalmente em transmitir a mensagem certa ou pode falar muito. Os dados e os visuais precisam trabalhar juntos, e é uma arte combinar uma ótima análise com uma ótima narrativa.

Edward Tufte, em seu livro de 1983, “The Visual Display of Quantitative Information”, explicou que os usuários de exibições de informações estão executando tarefas analíticas específicas, como fazer comparações. O princípio de design do infográfico deve apoiar a tarefa analítica. Segundo William Cleveland e Robert McGill, diferentes elementos gráficos realizam isso de forma mais ou menos eficaz. Por exemplo, gráficos de pontos e gráficos de barras superam os gráficos de setor (pizza).

Ainda na obra “The Visual Display of Quantitative Information”, Edward Tufte define exibições gráficas e princípios para exibição gráfica eficaz na seguinte passagem: "Excelência em gráficos estatísticos consiste em ideias complexas comunicadas com clareza, precisão e eficiência. As exibições gráficas devem:

- mostrar os dados;
- levar o espectador a pensar sobre a substância em vez de metodologia, design gráfico, tecnologia de produção gráfica ou qualquer outra coisa;
- evitar distorcer o que os dados têm a dizer;
- apresentar muitos números em um espaço pequeno;
- tornar grandes conjuntos de dados coerentes;

- encorajar o olho a comparar diferentes pedaços de dados;
- revelar os dados em vários níveis de detalhe, desde uma visão geral ampla até a estrutura fina;
- servir a um propósito razoavelmente claro: descrição, exploração, tabulação ou decoração;
- ser estreitamente integrado com as descrições estatísticas e verbais de um conjunto de dados.

Gráficos revelam dados. Na verdade, os gráficos podem ser mais precisos e reveladores do que os cálculos estatísticos convencionais".

Por exemplo, o diagrama de Minard mostra as perdas sofridas pelo exército de Napoleão no período de 1812-1813. Seis variáveis são plotadas: o tamanho do exército, sua localização em uma superfície bidimensional (x e y), o tempo, a direção do movimento e a temperatura. A largura da linha ilustra uma comparação (tamanho do exército em pontos no tempo), enquanto o eixo da temperatura sugere uma causa da mudança no tamanho do exército. Essa exibição multivariada em uma superfície bidimensional conta uma história que pode ser compreendida imediatamente ao identificar os dados de origem para criar credibilidade. Tufte cita, em 1983, que: "Pode muito bem ser o melhor gráfico estatístico já desenhado".

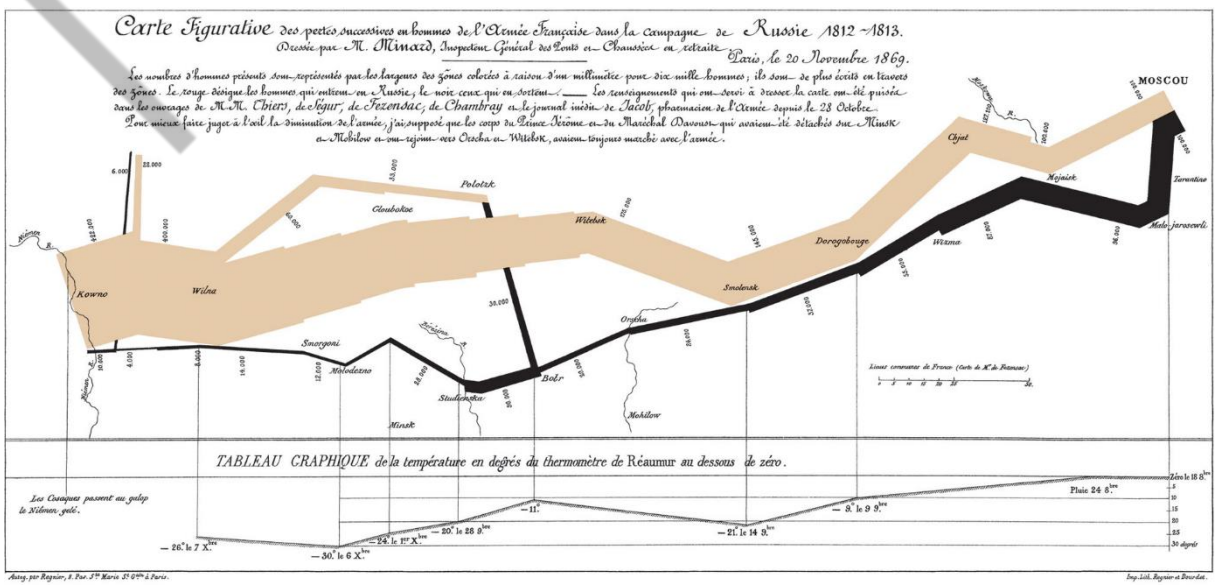


Figura 1 – Diagrama de Minard de 1869
Fonte: Revista Imagem FSG (Faculdade da Serra Gaúcha) (2015)

A não aplicação desses princípios pode resultar em gráficos enganosos, distorcer a mensagem ou apoiar uma conclusão errônea. De acordo com Tufte, chartjunk refere-se à decoração interior estranha do gráfico que não aprimora a mensagem ou efeitos tridimensionais ou de perspectiva gratuitos. Separar desnecessariamente a chave explicativa da própria imagem, exigindo que o olho viaje da imagem para a chave, é uma forma de "resíduos administrativos". A proporção de "dados para tinta" deve ser maximizada, apagando tinta que não seja de dados sempre que possível.

O Congressional Budget Office resumiu várias práticas recomendadas para exibições gráficas em uma apresentação de junho de 2014. Estes incluíram:

1. Conhecer o seu público.
2. Desenhar gráficos que possam estar isolados fora do contexto do relatório.
3. Criação de gráficos que comuniquem as principais mensagens do relatório.

O QUE VOCÊ VIU NESTA AULA?

Nesta aula vimos como iniciar no mundo das visualizações gráficas, editando parâmetros importantes que nos ajudam a visualizar dados para que no futuro eles possam gerar indicadores do problema em questão. Além disso, foi possível identificar com embasamento teórico, o motivo pelo qual o desenvolvimento de uma boa visualização pode ser chave.

Daqui para a frente, é importante que você replique os conhecimentos adquiridos para fortalecer ainda mais as suas bases e conhecimentos, já que um bom ou uma boa analista de dados de dados não é aquele(a) que é uma enciclopédia humana, mas sim aquele(a) que sabe ler um problema e atuar com eficácia.

IMPORTANTE: não esqueça de acessar o documento que contém o desafio da disciplina, para que assim você possa aprimorar os seus conhecimentos de maneira sólida, além, é claro, de interagir no Discord e participar das lives!

REFERÊNCIAS

CLEVELAND, WS; MCGILL, R. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association, vol. 79, nu. 387, p. 531-554, set 1984.

DOCUMENTAÇÃO MATPLOTLIB. **<https://matplotlib.org/>.** Acesso em 06 fev. 2023.

DOCUMENTAÇÃO PANDAS. **<https://pandas.pydata.org/>.** Acesso em: 06 fev. 2023.

GOOGLE COLAB. **<https://colab.research.google.com/>.** Acesso em: 06 fev. 2023.

PISSETTI, Rodrigo Fernandes. Princípios fundamentais do design analítico. Revista Imagem v, v. 5, n. 1. Rio Grande do Sul, 2015.

SHIFT LEARNING. <**https://www.shiftelearning.com/blog/bid/350326/studies-confirm-the-power-of-visuals-in-elearning**>. Acesso em: 07 fev 2023.

TUFTE, Edward R. The visual display of quantitative information. Cheshire, Connecticut: Graphics Press, 1983.

PALAVRAS-CHAVE

Palavras-chave: Python. Pandas. Dataframe.