Aula 02 | PosTech | Análise Exploratória de Dados

Anotações sobre a segunda aula da PosTech FIAP ✨✨

https://on.fiap.com.br/mod/conteudoshtml/view.php?id=307785&c=8729&sesskey=vCMoHFxpWh

Temas abordados:

- Visualização de dados do Dataframe utilizando Matplotlib;
- Editar parâmetros importantes para visualizar melhor os dados;
- Saber lidar com os dados que temos em mãos.

Dependências:

- Baixar o arquivo zip:

https://github.com/alura-tech/pos-datascience-analise-e-exploracao-de-dados/tree/aula1

- Documentação Pandas:

https://pandas.pydata.org/

- Documentação Matplotlib:

https://matplotlib.org/

- Google Colab:

https://colab.research.google.com/

- TabNet:

https://datasus.saude.gov.br/informacoes-de-saude-tabnet/

```
import matplotlib.ticker as ticker
import matplotlib.pyplot as plt

axis = dados.plot(x='Unidade da Federação', y='2008/Ago', kind="bar", figsize=(9,6))
axis.yaxis.set_major_formatter(ticker.StrMethodFormatter('{x:,.2f}'))

plt.title('Valor por unidade da federação')
plt.show()
```

Aula 2 - Primeiras Visualizações de Dados

Biblioteca Pandas → Uma das mais completas quando o assunto é análise de dados, porque nos permite *ler e gravar arquivos* em diferentes 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).

Biblioteca Matplotlib → A mais usada para *visualização de dados em forma de gráfico*, além de ser base para a criação de outras bibliotecas (Seaborn e Plotly). 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…

1 | DATAVIZ → “Data Visualization”

Em um mundo orientado a dados, é importante uma **boa visualização** para a execução de **conclusões de qualidade**.

Data Visualization é a forma pela qual representamos nosso conjunto de dados, sejam eles **estruturados** ou **não**. 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.

É amplamente utilizada em muitas áreas, incluindo **negócios, ciências, saúde, tecnologia e governo, para ajudar a tomar decisões informadas** e a compreender **informações complexas**.

Nosso cérebro é muito **mais eficiente em processar informações visuais** do que dados brutos. É 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.

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 menor de lembrá-los no futuro. (!) Constatou-se que, após três dias, os estudos analisados retinham entre 10% e 20% das informações escritas ou faladas, em comparação com 65% das informações visuais.

Há muitas técnicas e ferramentas disponíveis para ajudar na visualização de dados, incluindo:

gráfico de barras,

gráfico de linhas,

gráficos de dispersão,

mapas

e infográficos.

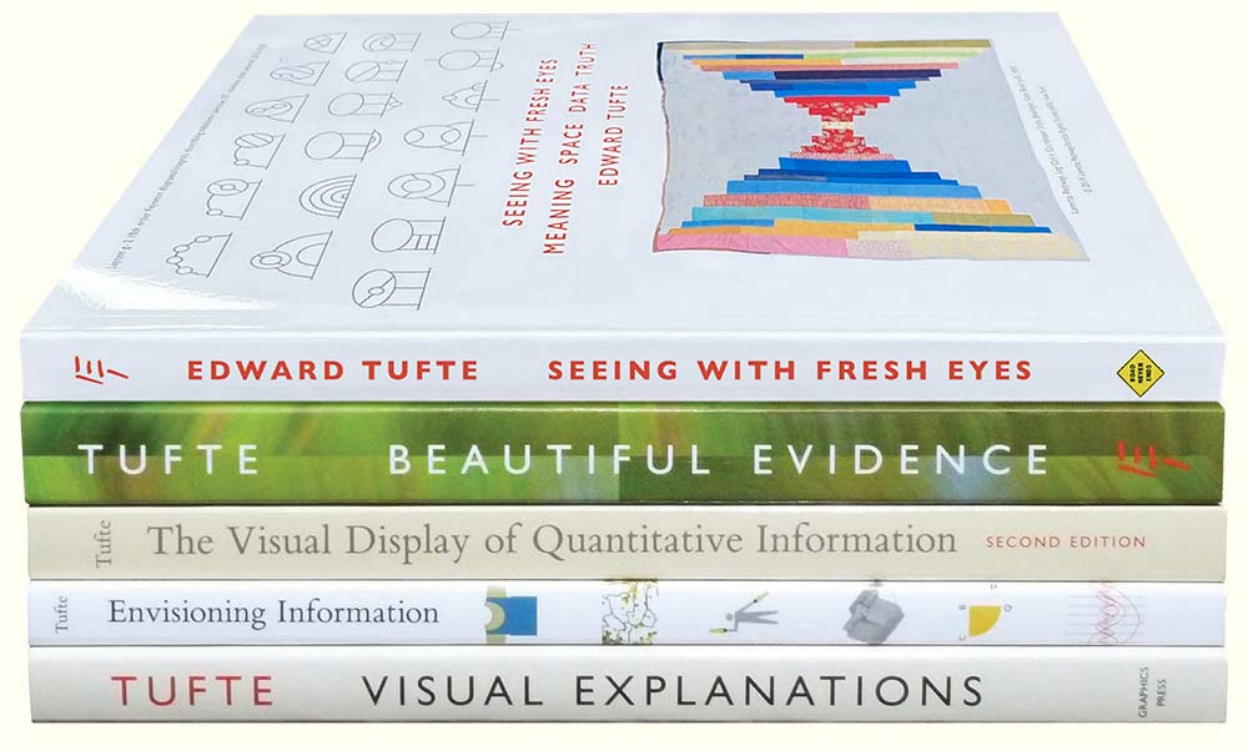
Cada tipo de visualização é adequado para diferentes tipos de dados e objetivos, é importante escolher a visualização adequada para garantir que a informação seja transmitida de forma clara e precisa.

Visualização de dados → comunicação de informações de maneira clara e memorável para públicos de diferentes contextos e níveis de conhecimento técnico ✨✨✨

→ útil em apresentações **comerciais**, onde é necessário convencer outras pessoas a tomar uma decisão ou adotar uma perspectiva específica.

→ "Era do Big Data" → visualização é ferramenta cada vez mais importante para entender os trilhões de linhas de dados gerados todos os dias → ajuda a contar histórias, organizando dados de forma fácil de entender, destacando tendências e valores discrepantes → boa visualização conta uma história, removendo o ruído dos dados e destacando informações úteis.

⚠️ Visualização efetiva de dados é equilíbrio entre FORMA e FUNÇÃO. Dados e visuais precisam trabalhar juntos e é uma arte combinar uma *ótima análise com uma ótima narrativa*. ⚠️



Edward Tufte (1983) → "**The Visual Display of Quantative Information**" → explicou que os usuários de exibições de informações estão executando tarefas analíticas específicas, como fazer comparações. O princípio do 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).

→ gráficos de pontos | gráficos de barras **>>>>>>** gráficos de setor (pizza)

Edward, em sua obra, define exibições gráficas e princípios para exibição gráfica eficaz na passagem:

“Excelência em gráficos estatísticos consiste em ideias complexas comunicadas com ***clareza, precisão e eficiência****. As exibições gráficas devem:*

- *Mostrar dados;*
- *Levar o espectador a pensar sobre a substância em vez da 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.”

Exemplo: Diagrama de Minard → mostra as perdas sofridas pelo exército de Napoleão no período de 1812-1813.

Seis variáveis são plotadas:

1 | tamanho do exército;

2 | localização em superfície bidimensional (x,y)

3 | tempo

4 | direção do movimento

5 | temperatura

A largura da linha ilustra uma comparação (tamanho do exército em pontos no tempo), enquanto o eixo da temperatura sugere uma causa na mudança do 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.

Edward cita, em 1983, que: “Pode muito bem ser o melhor gráfico estatístico já desenhado”.

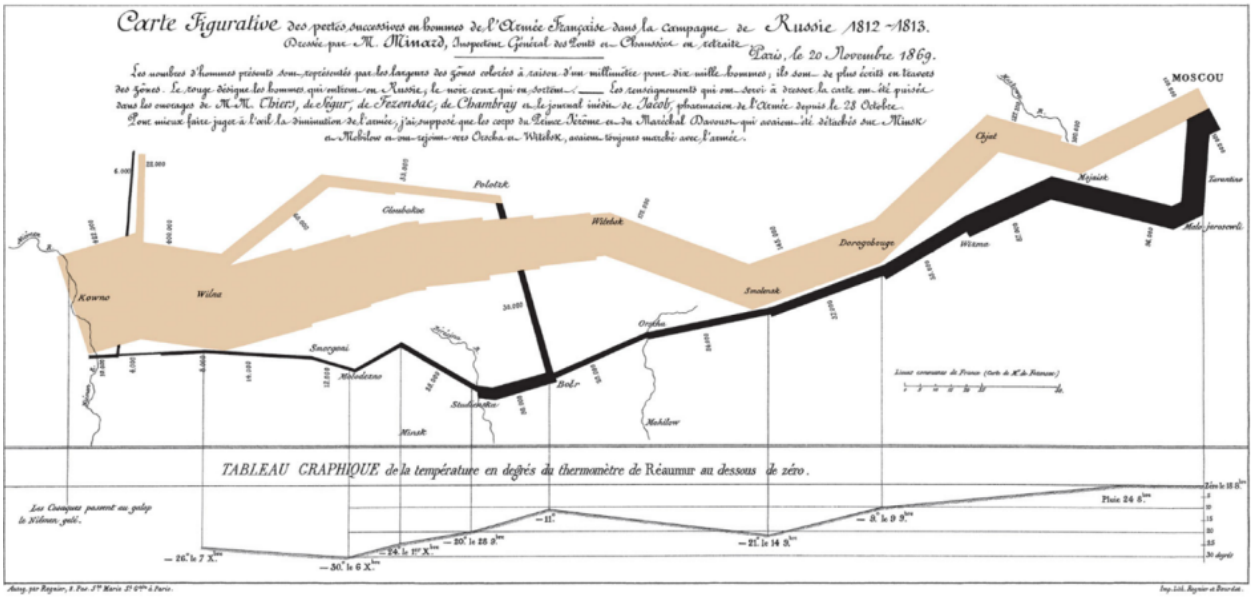


Diagrama de Minard (1869)

As consequências da não aplicação desses princípios pode resultar em:

1 | gráficos enganosos

2 | distorcer a mensagem

3 | apoiar uma conclusão errônea

De acordo com Edward, 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. Incluíram:

1 | Conhecer bem 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.

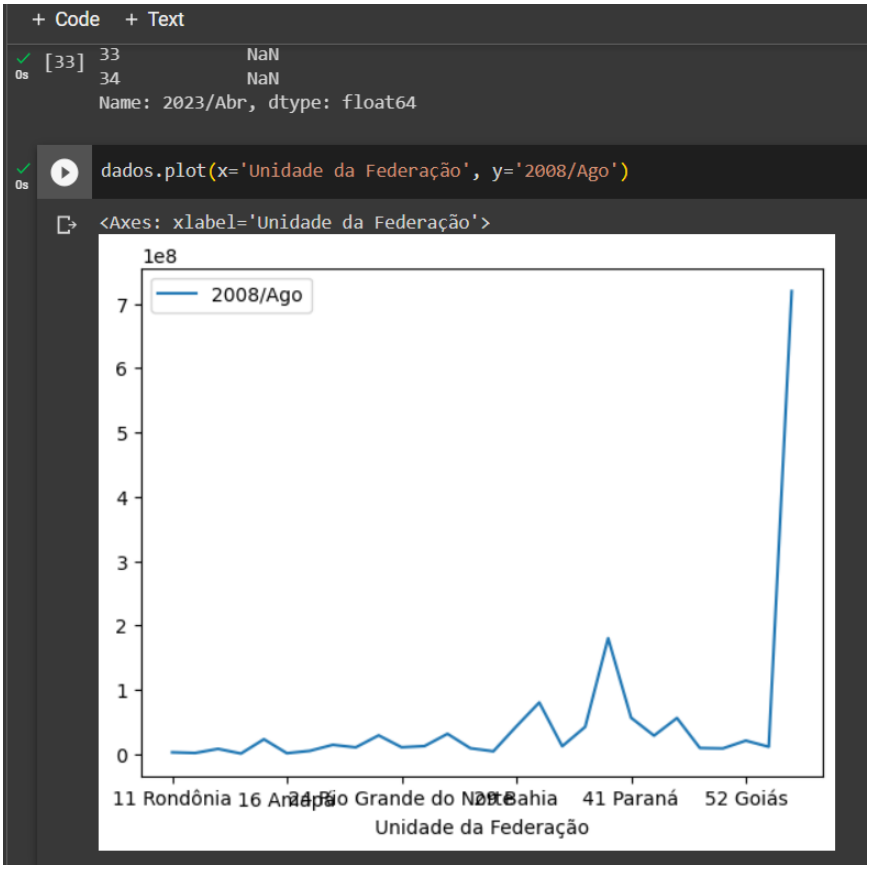
2 | AULA PRÁTICA | 01 DE 01

Considerando onde paramos na aula anterior com base nos dados de Produção Hospitalar, agora queremos PLOTAR alguns dados.

função plot

- [ ] `dados.plot(x=”Unidade da Federação”, y=”2008/Ago”)`

Quero plotar valores no mês de Agosto de 2008. No eixo x quero a “Unidade da Federação” e no y eu quero uma única coluna, que é a de “2008/Ago”.



Aqui, temos um “**gráfico de linha**” que começa com “11 Rondônia” e vai até “52 Goiás”. No eixo y vai de 0 até 7 milhões em 2008/Ago. Então, a gente tem um gráfico de linha que é plotado. Com dados.plot a gente consegue diversos tipos de visualizações através de gráficos. 🙂

Esse gráfico, na verdade, é um horror. HAHAHAHA

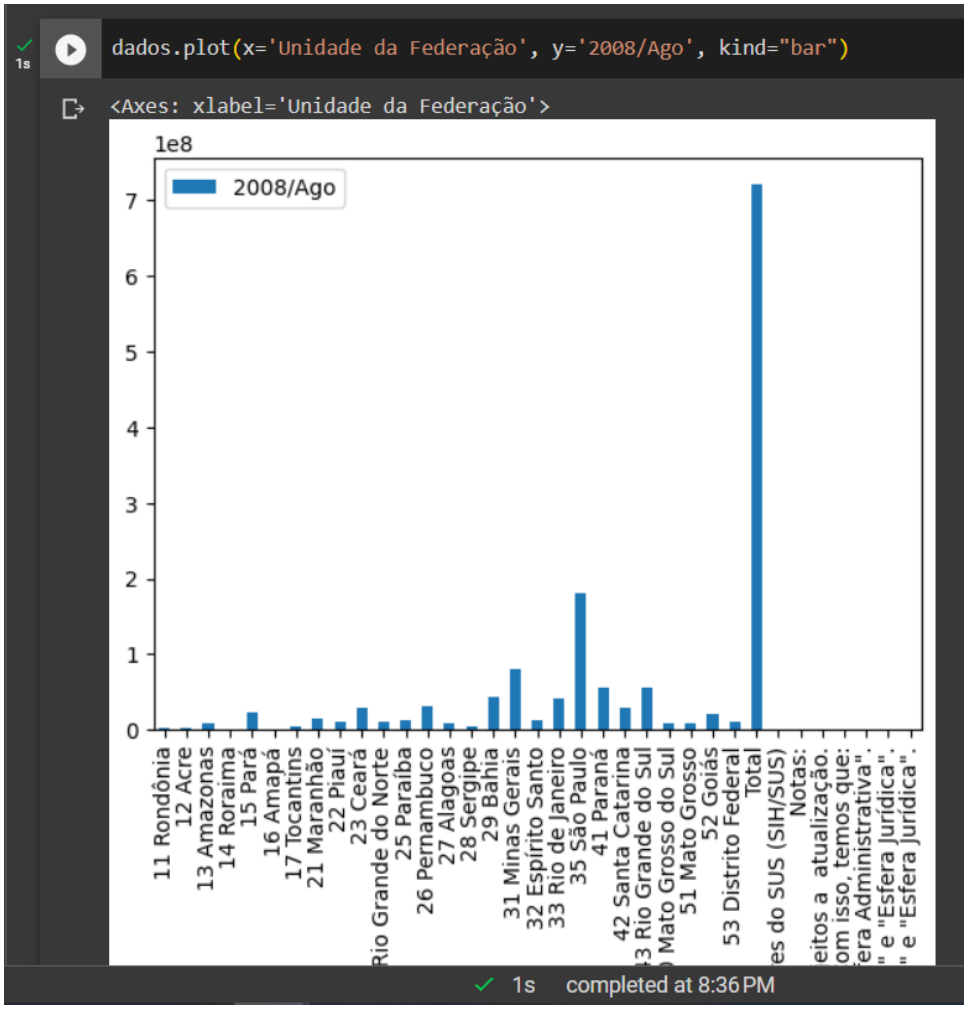
Não faz sentido um gráfico de linhas para dizer valores de estados. Se eu conecto dois pontos, significa que existe alguma coisa entre esses dois pontos. No caso de um estado isso não faz o menor sentido. Existe algum estado entre o Paraná e o Goiás? Existe. Os estados que estão entre o Paraná e Goiás são estados que realmente ficam fisicamente ficam entre Paraná em Goiás? Não faz sentido isso.

Precisamos saber escolher um gráfico adequado para saber passar a representação ideal. 😊

estilizando o gráfico

E se eu não quiser mais um gráfico de linha mas sim um gráfico de barra?

- [ ] `dados.plot(x=”Unidade da Federação”, y=”2008/Ago”, kind="bar")`

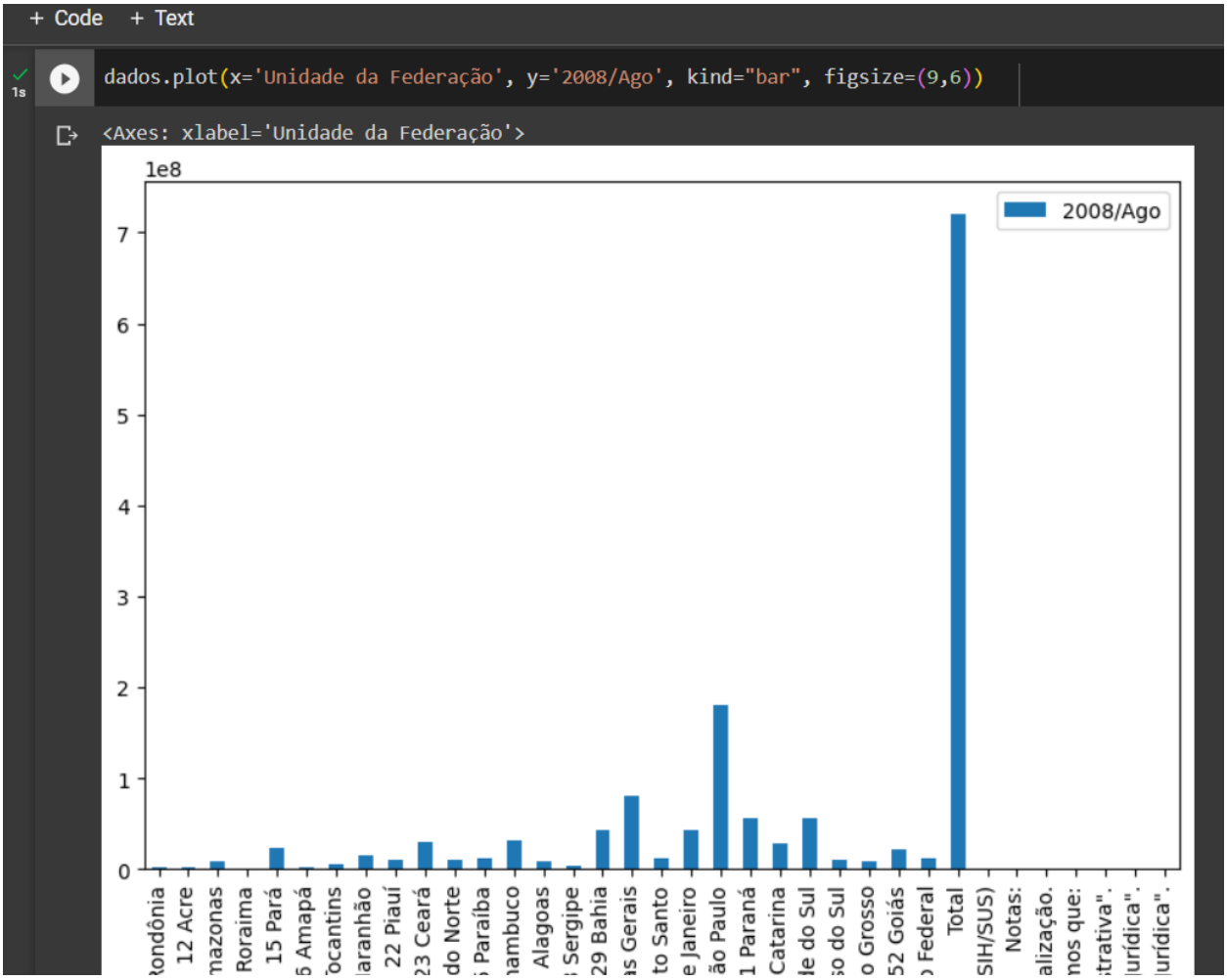


Agora sim tenho todos os estados, todas as unidades da federação.

São Paulo teve o maior gasto nesse mês. Comparativamente, MG com a metade. No gráfico anterior, não conseguira saber nem onde está SP ou MG. O gráfico, portanto, conta uma mensagem que eu e você queremos passar. Por isso, desconfie de todos os gráficos a partir de agora. HAHAHAHAHA É uma mensagem de um ser humano e pode sempre ter algum tipo de viés sendo passado sem querer.

Mas achei muito pequeno…

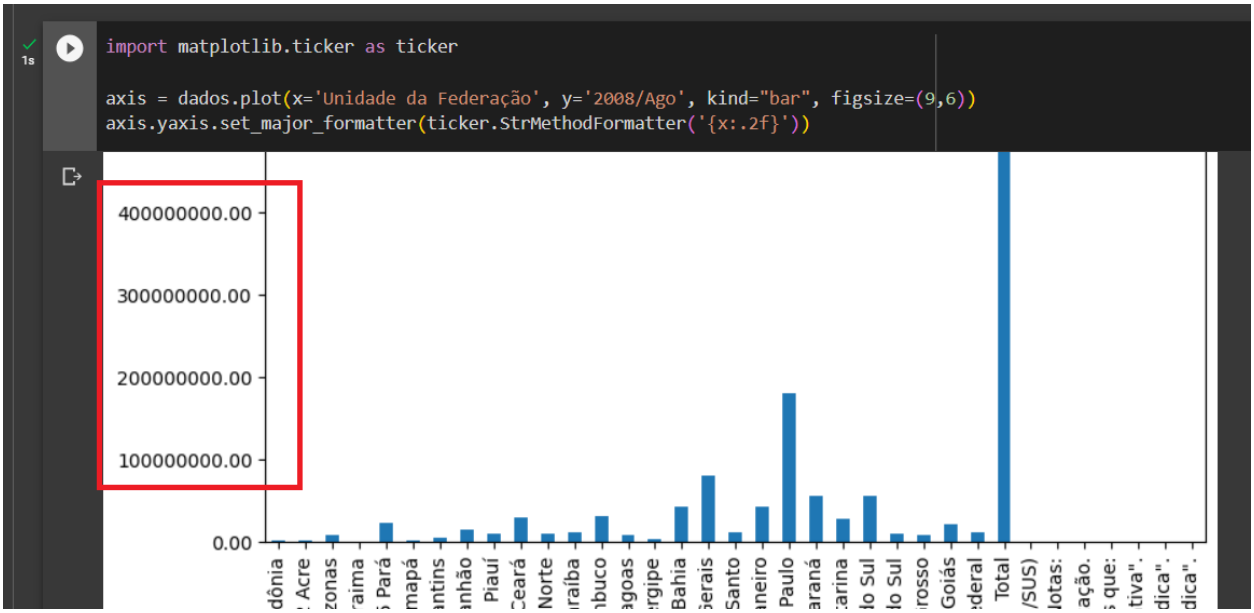
```
dados.plot(x=”Unidade da Federação”, y=”2008/Ago”, kind="bar", figsize=(9,6))
```



Ficou melhor. 😀

O panda devolve algo chamado “axis”, que é “eixo”. Mas não quero todos os eixos, quero apenas o eixo y.

- [ ] `import matplotlib.ticker as ticker`
- [ ] `axis = dados.plot(x='Unidade da Federação', y='2008/Ago', kind="bar", figsize=(9,6))`
- [ ] `axis.yaxis.set_major_formatter(ticker.StrMethodFormatter(”{x:.2f}”))`



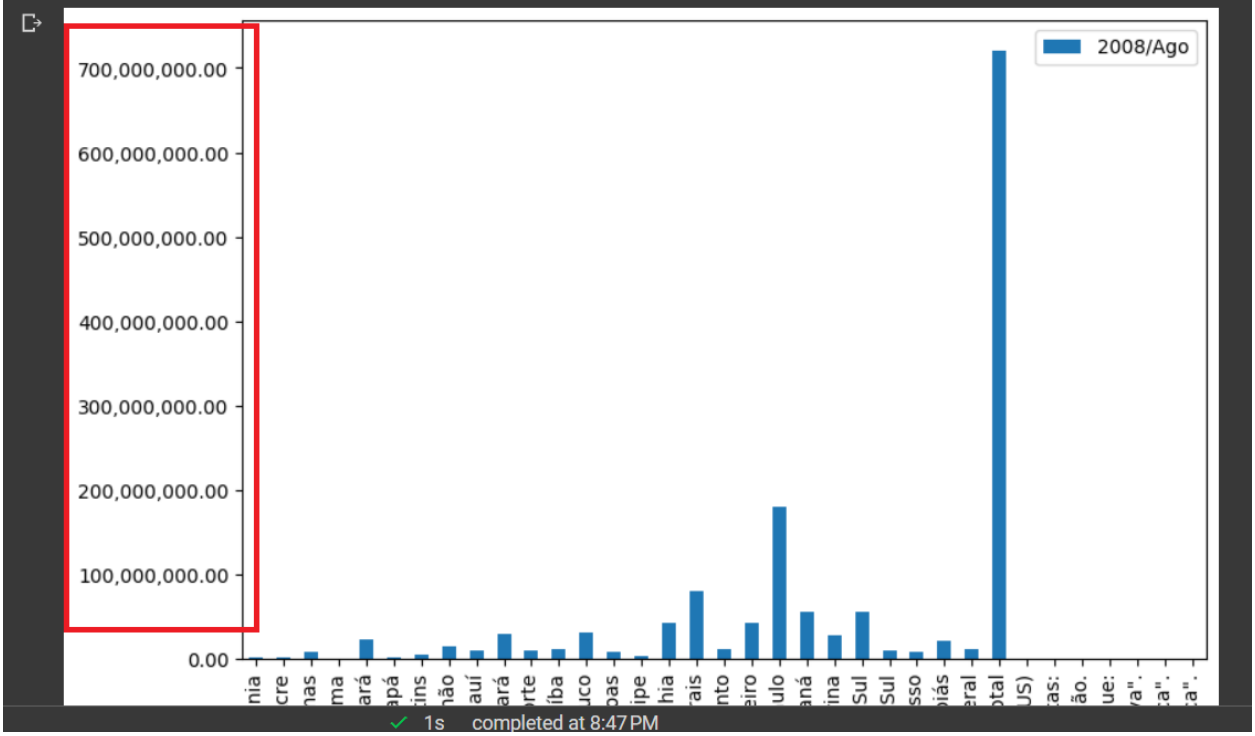
Tick formatters — Matplotlib 3.1.2 documentation

https://matplotlib.org/3.1.1/gallery/ticks_and_spines/tick-formatters.html

- [ ] `axis.yaxis.set_major_formatter(ticker.StrMethodFormatter(”{x:,.2f}”))`

```
import matplotlib.ticker as ticker

axis = dados.plot(x='Unidade da Federação', y='2008/Ago', kind="bar", figsize=(9,6))
axis.yaxis.set_major_formatter(ticker.StrMethodFormatter('{x:,.2f}'))
```



1s completed at 8:47 PM

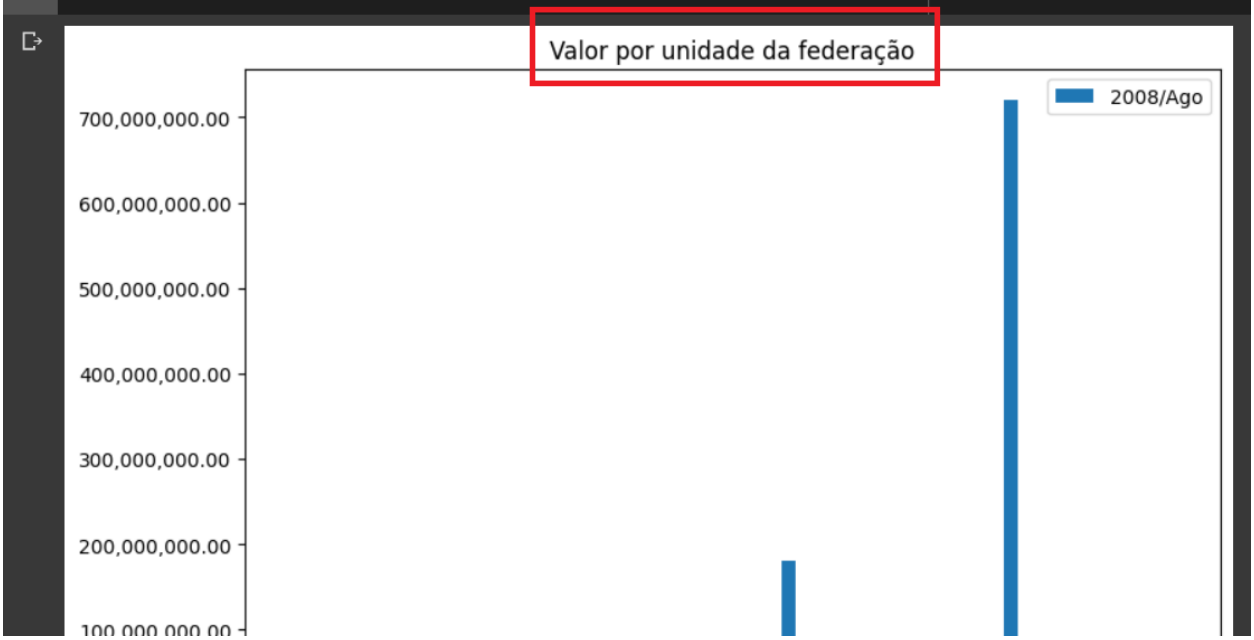
Além disso, quero colocar um título. O título é no grande plot, não no eixo.

- [ ] `import matplotlib.pyplot as plt`
- [ ] `plt.title(”Valor por unidade da federação”)`
- [ ] `plt.show()`

```
import matplotlib.ticker as ticker
import matplotlib.pyplot as plt

axis = dados.plot(x='Unidade da Federação', y='2008/Ago', kind="bar", figsize=(9,6))
axis.yaxis.set_major_formatter(ticker.StrMethodFormatter('{x:,.2f}'))

plt.title('Valor por unidade da federação')
plt.show()
```



Desafio 01: Fazer atualização desses valores para o mês mais recente que tenho acesso.

Desafio 02: Colunas na horizontal, não na vertical. Ou até mesmo 45º ou 30º, porque fica mais fácil de fazer a leitura dessa legenda. Dica: ler a documentação do Pandas.