Introdução a Ciência de Dado

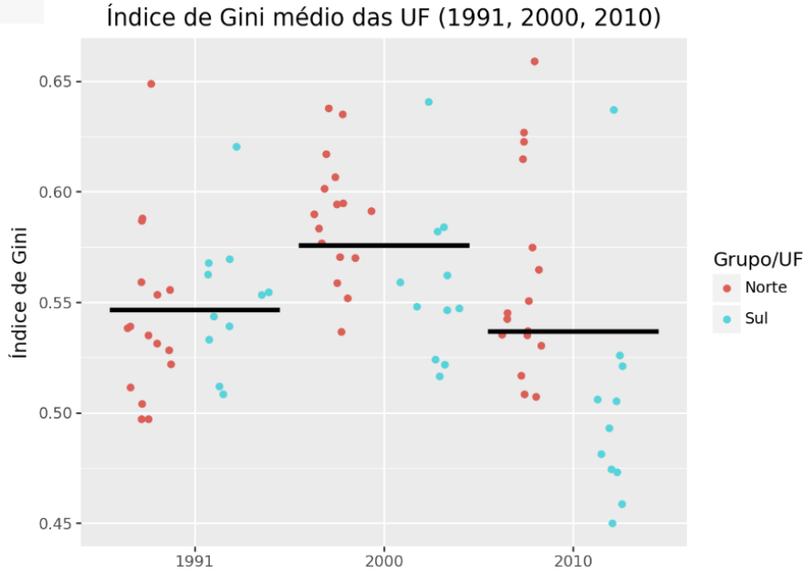
Professor: Alex Pereira

Comentários sobre o Exercício 0.1

- Como e porque despivotar uma tabela?

	Município	1991	2000	2010	cod_mun	codUF	UF	grupo
0	110001 Alta Floresta D'Oeste	0.5983	0.5868	0.5893	110001	11	RO	Norte
1	110037 Alto Alegre dos Parecis	NaN	0.5080	0.5491	110037	11	RO	Norte
2	110040 Alto Paraíso	NaN	0.6256	0.5417	110040	11	RO	Norte
3	110034 Alvorada D'Oeste	0.5690	0.6534	0.5355	110034	11	RO	Norte
4	110002 Ariquemes	0.5827	0.5927	0.5496	110002	11	RO	Norte

	Município	cod_ibge	uf	grupo	ano	gini
0	110001 Alta Floresta D'Oeste	110001	11	Norte	1991	0.5983
3	110034 Alvorada D'Oeste	110034	11	Norte	1991	0.5690
4	110002 Ariquemes	110002	11	Norte	1991	0.5827
6	110003 Cabixi	110003	11	Norte	1991	0.6527
8	110004 Cacoal	110004	11	Norte	1991	0.6800



Comentários sobre o Exercício 0.1

- Como despivotar uma tabela?

```
df_long = pd.melt(df, id_vars=['Município', 'cod_ibge', 'uf', 'grupo’],
                      value_vars=['1991', '2000', '2010’],
                      var_name='ano', value_name='gini')
```

	Município	cod_ibge	uf	grupo	ano	gini
0	110001 Alta Floresta D'Oeste	110001	11	Norte	1991	0.5983
3	110034 Alvorada D'Oeste	110034	11	Norte	1991	0.5690
4	110002 Ariquemes	110002	11	Norte	1991	0.5827
6	110003 Cabixi	110003	11	Norte	1991	0.6527
8	110004 Cacoal	110004	11	Norte	1991	0.6800

Como criar uma coluna categórica baseado num mapeamento

```
norte_nordeste = [
    '11','12','13','14','15','16','17',  # Norte
    '21','22','23','24','25','26','27','28','29'  # Nordeste
]
df['grupo'] = np.where(df['uf'].isin(norte_nordeste), 'Norte', 'Sul')
```

Comentários sobre o Exercício 1.1

- Sempre prestar atenção ao significado exato dos valores plotados nos eixos
  - E explicar para a sua audiencia, quando não for intuitive
    - ✓ Se importar em saber o significado desses valores
      - Média global de temperatura? Onde foi coletado? Quantos pontos de coleta? No oceano? Na terra? A que altura?
- Como foi mapeado os valores numéricos do eixo x nos respectivos meses?
- Utilidade do intervalo de confiança

Comentários sobre o Exercício 1.1

```
baseline_tmax = [15.08, 15.34, 16.18, 17.33, 18.52, 19.40, 19.78, 19.59, 18.90,
17.63, 16.26, 15.33]


baseline_df = pd.DataFrame({"Month": range(1, 13), "Baseline": baseline_tmax})


df = df.merge(baseline_df, on="Month", how="left")


df["AbsoluteTemp"] = df["MonthlyAnomaly"] + df["Baseline"]
```

Comentários sobre o Exercício 1.1

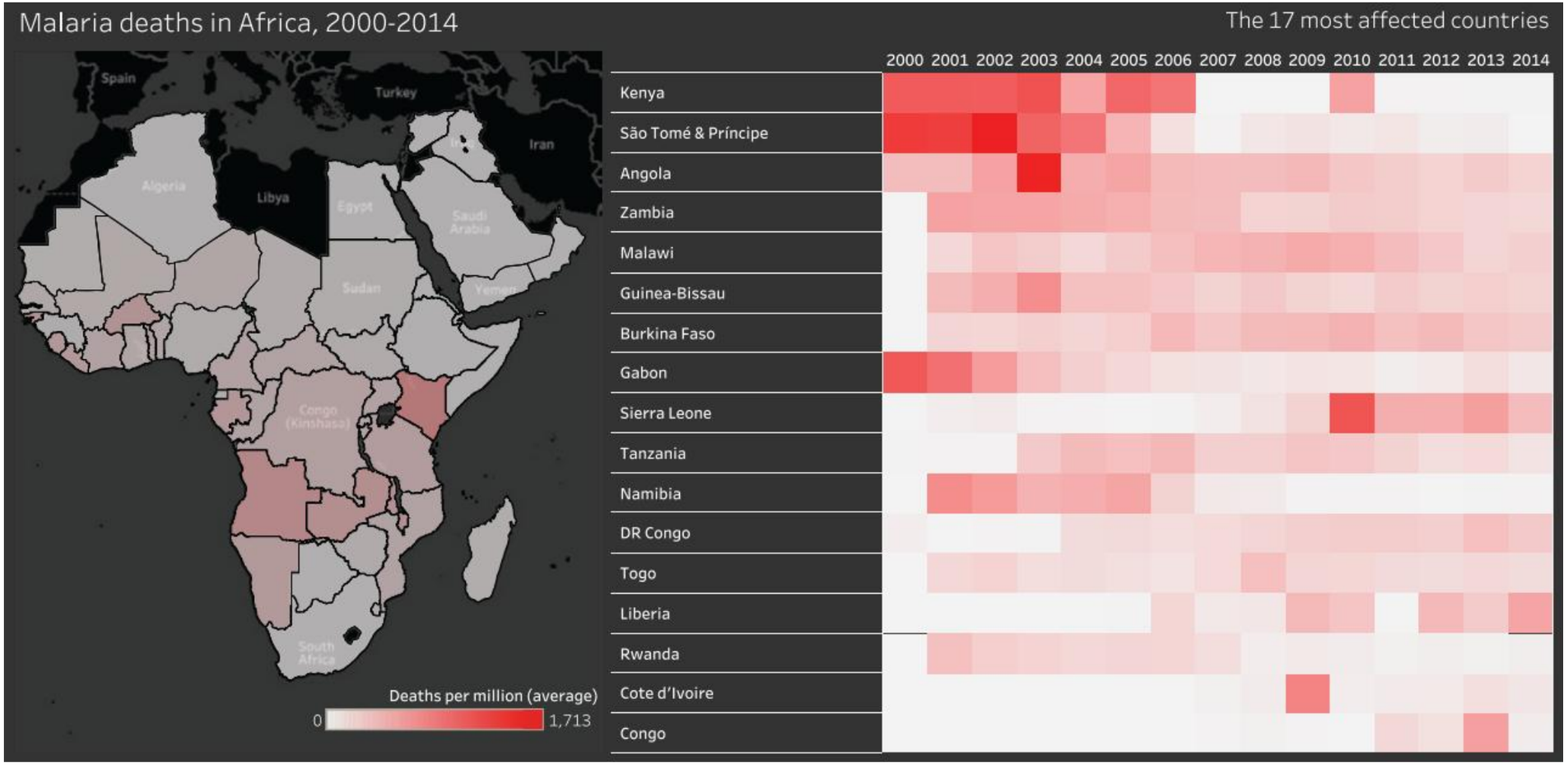
- Sempre prestar atenção ao significado exato dos valores plotados nos eixos
  - E explicar para a sua audiencia, quando não for intuitive
    - ✓ Se importar em saber o significado desses valores
      - Média global de temperatura? Onde foi coletado? Quantos pontos de coleta? No oceano? Na terra? A que altura?
- Como foi mapeado os valores numéricos do eixo x nos respectivos meses?
- Utilidade do intervalo de confiança
- Soluções de referência
  - Francisco, Giacomin e Alex
- Dashboard de Notas

Tipos de Dados

- Categórico
  - Rótulos que podem assumir um número limitado/fixo de opções
    - ✓ Profissão, escola, cidade, característica social/demográfica
- Ordinal
  - Semelhante ao categórico, porém com ordenação
    - ✓ Dias da semana, meses, postos/cargos de trabalho, aulas sucessivas
- Quantitativo
  - Valores numéricos discretos ou contínuos
    - ✓ Datas (quantitativo e ordinal),
    - ✓ Nota (métricas de desempenho), idade, pageviews

Codificação de informação em gráficos

- Classifique os tipos de informação apresentados em
  - Categórico, ordinal e quantitativo



Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Classificação dos dados do gráfico anterior

TABLE 1.5 Data used in the bar chart in Figure 1.14.

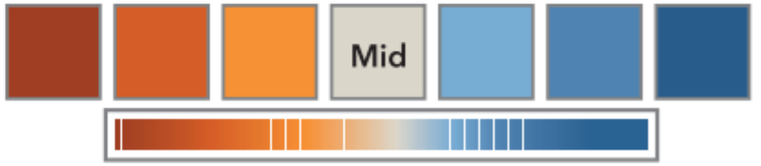
Data	Data Type	Encoding	Note
Country	Categorical	Position	The map shows the position of each country. In the highlight table, each country has its own row.
Deaths per million	Quantitative	Color	The map and table use the same color legend to show deaths per million people.
Year	Ordinal	Position	Each year is a discrete column in the table.

Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Recomendações de escalas de cores

SEQUENTIAL
color is ordered from low to high

DIVERGING
two sequential colors with a neutral midpoint



CATEGORICAL
contrasting colors for individual comparison

HIGHLIGHT
color used to highlight something

ALERT
color used to alert or warn reader

FIGURE 1.16 Use of color in data visualization.

Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Principais tipos de gráficos

- De barras
  - Usam o comprimento para representar medidas
    - ✓ Somos muito bons para reconhecer pequenas diferenças, quando há uma linha de base comum
      - O comprimento é um dos atributos de pré-atenção mais eficientes para processarmos.
  - São muito efetivos para comparar categorias
- Gráficos de linha
  - Usualmente mostram mudanças ao longo do tempo
    - ✓ A inclinação da reta mostra tendências
- Gráfico de pizza
  - Evite!
    - ✓ Difícil comparar categorias de tamanho semelhante

O que há de peculiar nesse gráfico?

Principais tipos de gráficos

- Tabelas
  - Úteis para mostrar valores exatos
- Gráfico de pontos
  - Compara valores em 2 dimensões
- Existem muitas derivações
  - desses tipos elementares

Fonte: https://meneg.at/

Principais tipos de gráficos

- Não misture posição, cor e forma no mesmo gráfico

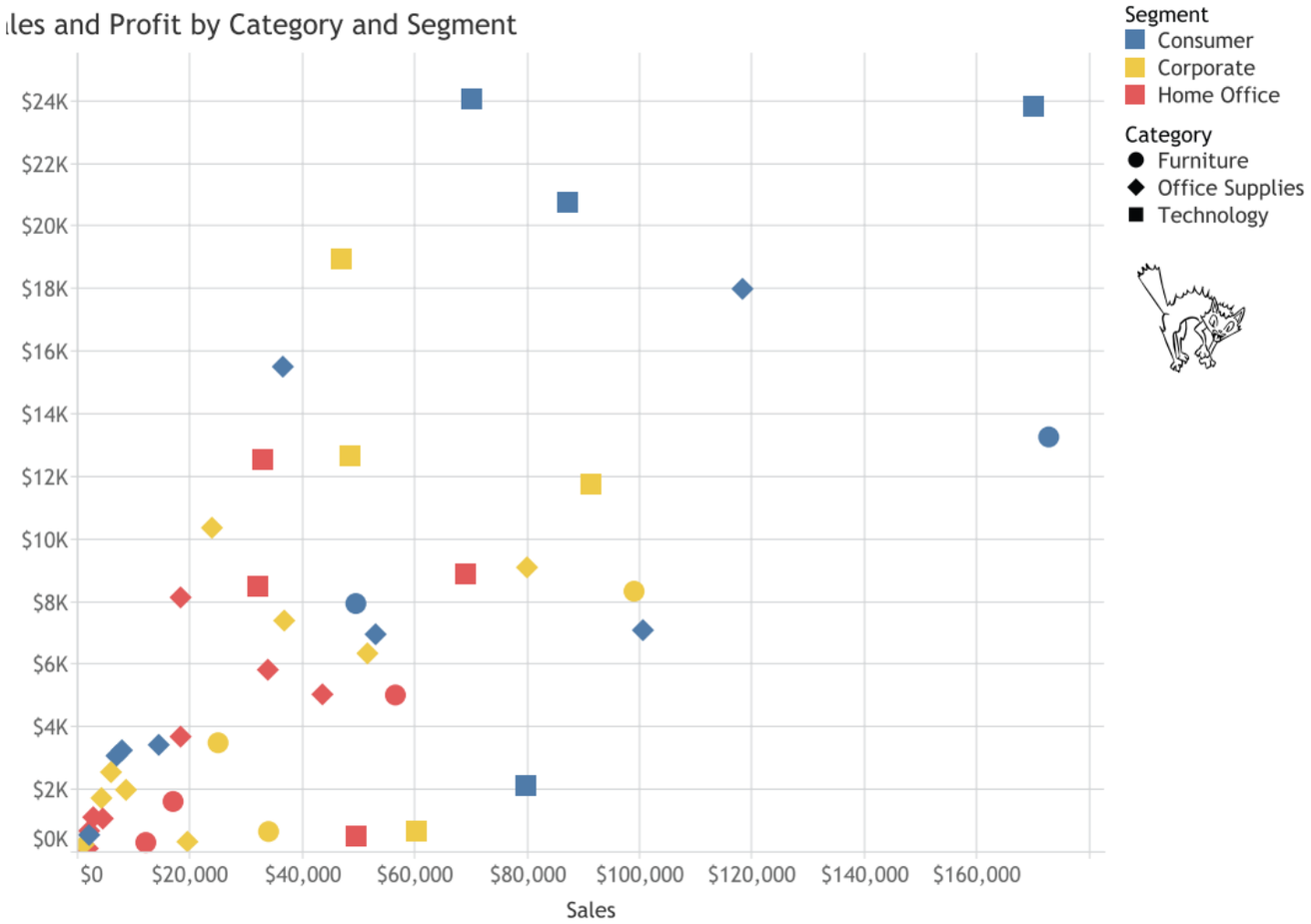


FIGURE 1.42 Scatterplot using shape and color. Which category has the highest profits?

Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Principais tipos de gráficos

- Posição representando as categorias
  - tecnologia, em média, tem lucratividade maior
    - ✓ do que Furniture (móveis) e Office Supplies (materiais de escritório).

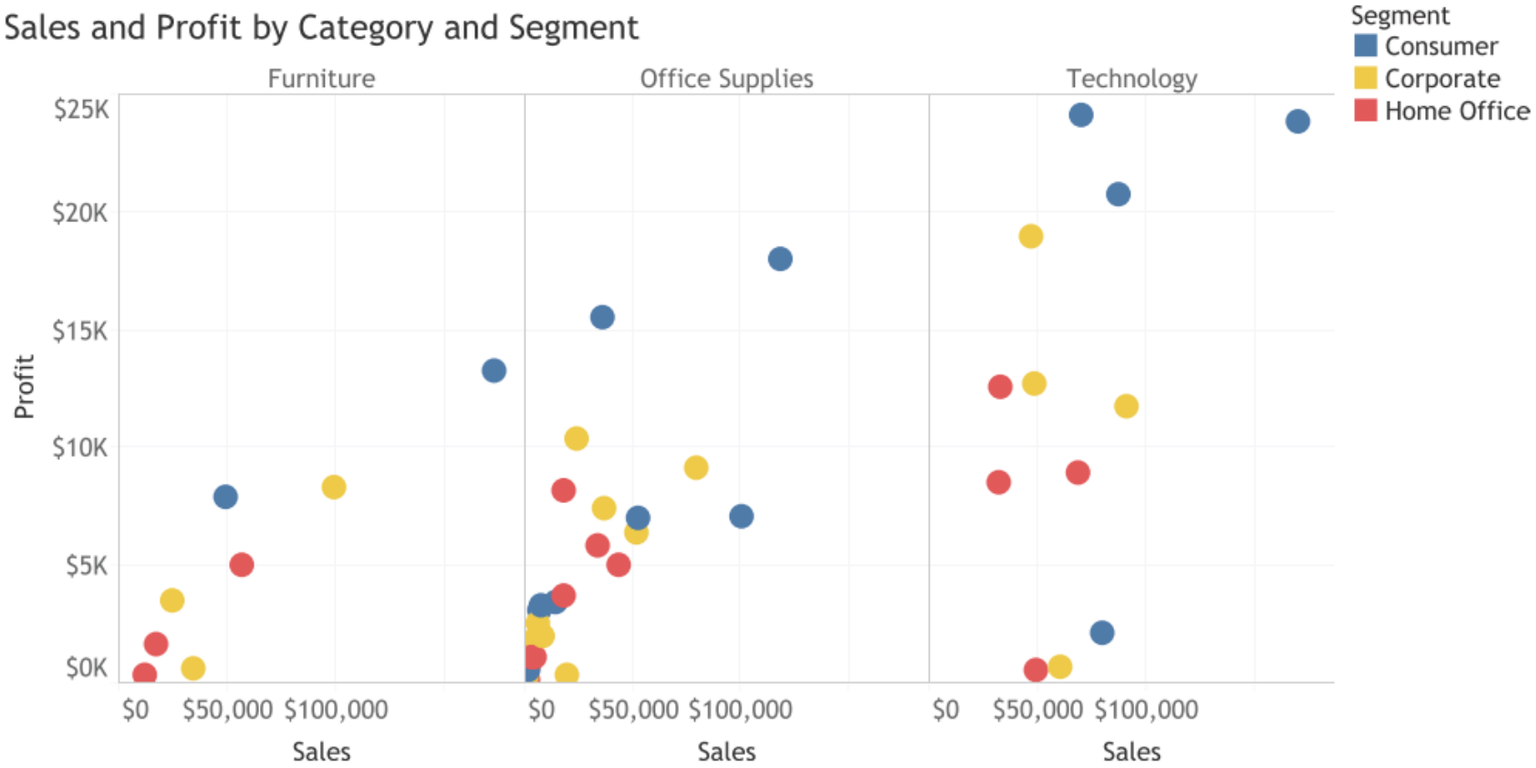


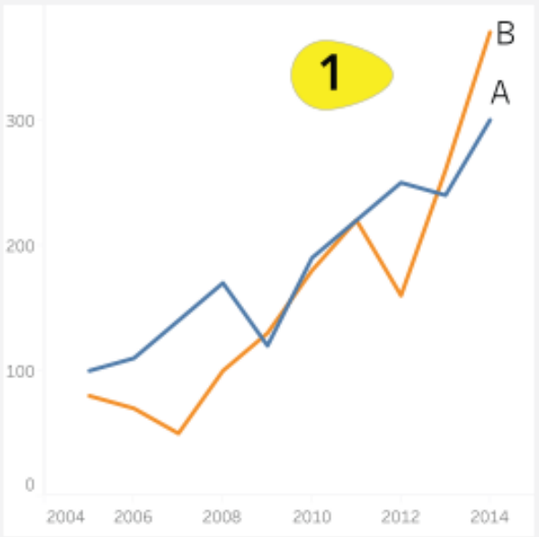
FIGURE 1.43 Sales and profit with one column for each category.

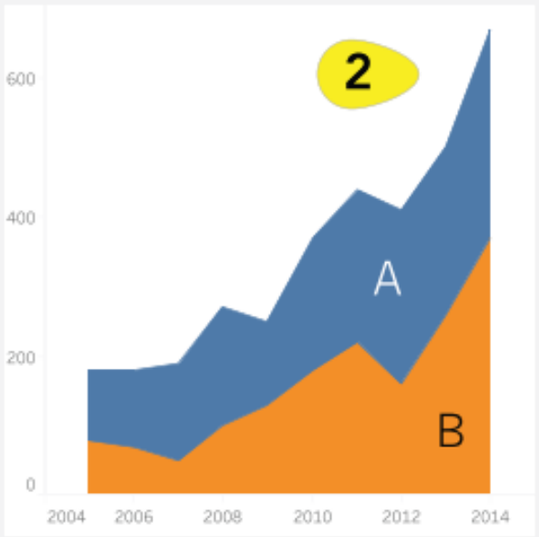
Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Qual gráfico é melhor ?

- Depende
  - Qual pergunta se quer responder ?
    - ✓ O gráfico 1 compara a venda de cada produto
    - ✓ O gráfico 2 mostra facilmente o total vendido ao longo do tempo

Year	A	B
2005	100	80
2006	110	70
2007	140	50
2008	170	100
2009	120	130
2010	190	180
2011	220	220
2012	250	160
2013	240	260
2014	300	370
Total	1,840	1,620





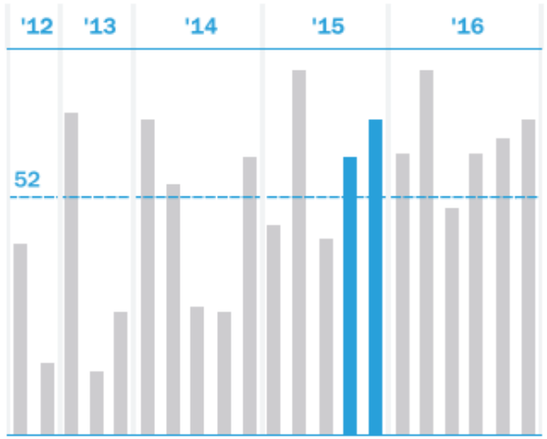
- Resposta pra quase tudo: depende da audiência
  - Quem são e seus objetivos.

Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Exemplo de Dashboard de uma Universidade

Course Metrics

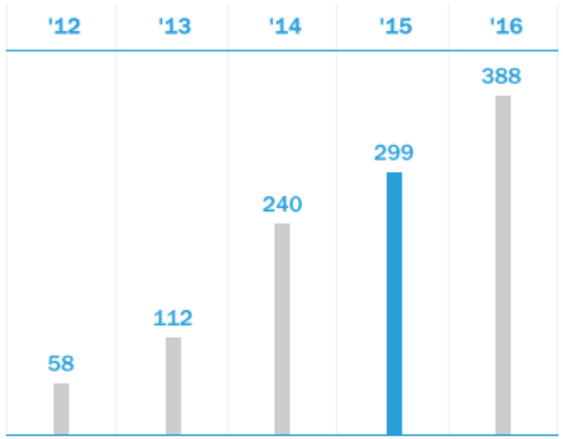
Students



1097

Total students in five years

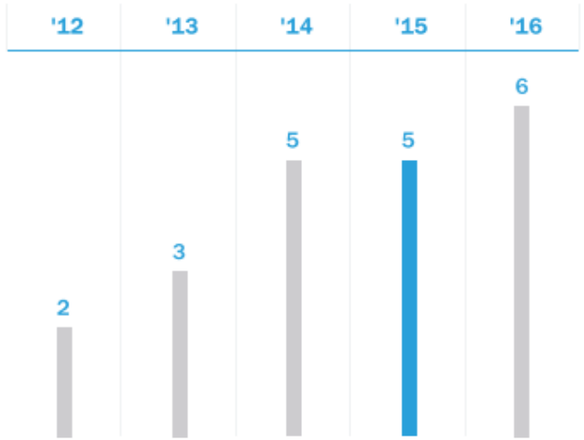
Enrollments



687

Total students in 2015-2016

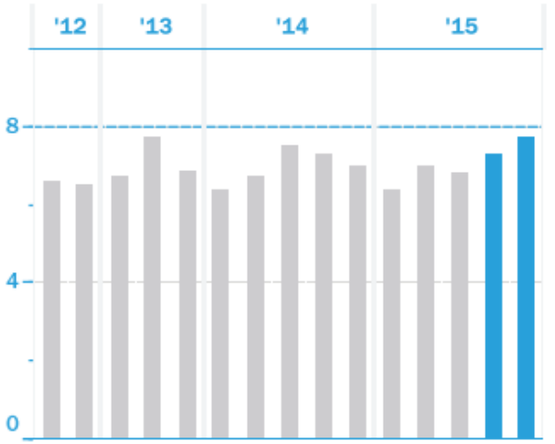
Classes



21

Total classes in five years

Ratings

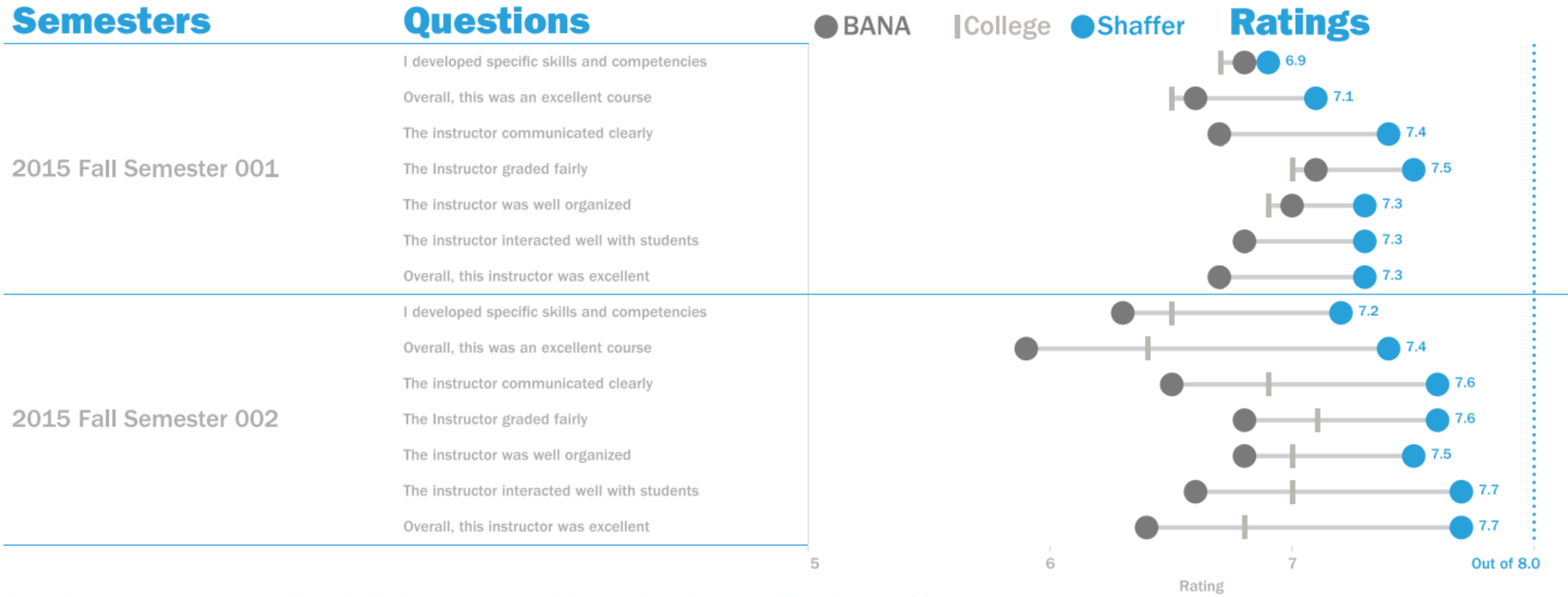


7.7 of 8

Most Recent Instructor Rating (out of 8.0)

Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Exemplo de Dashboard de uma Universidade

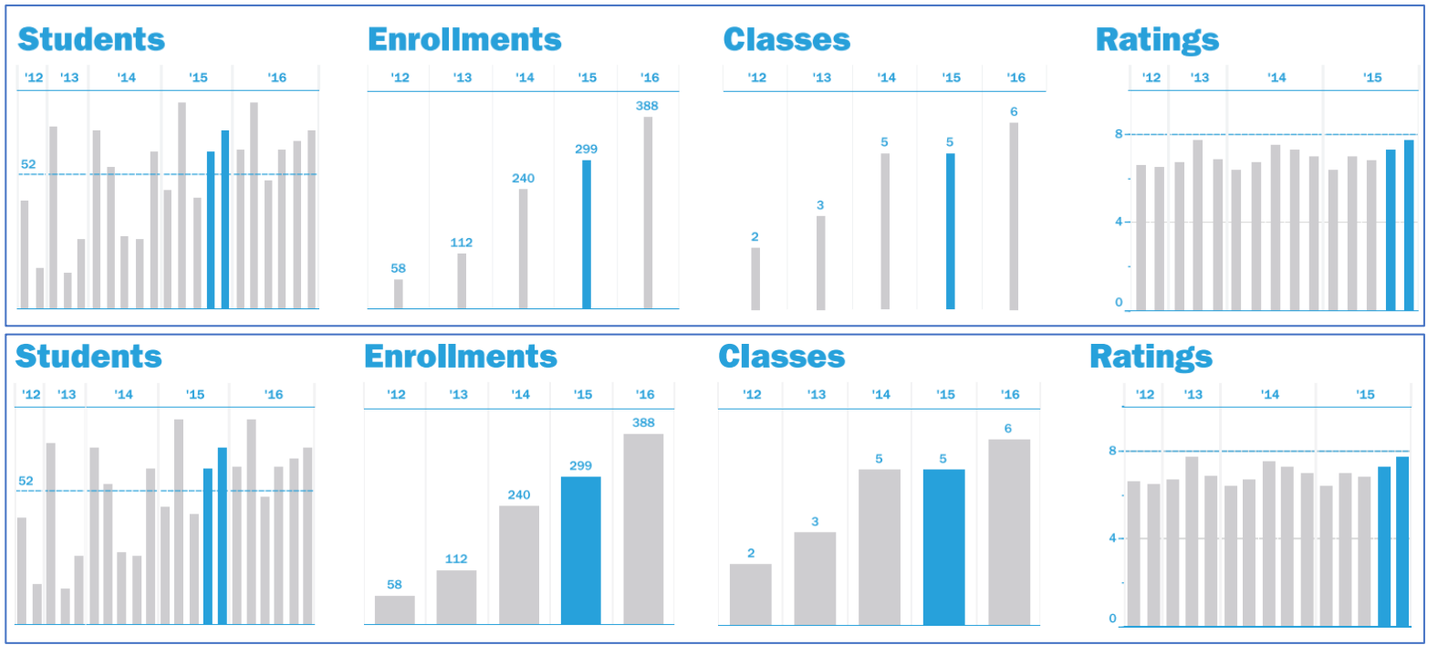


* BANA = Business Analytics; Shaffer = Professor do Curso de Visualização de Dados

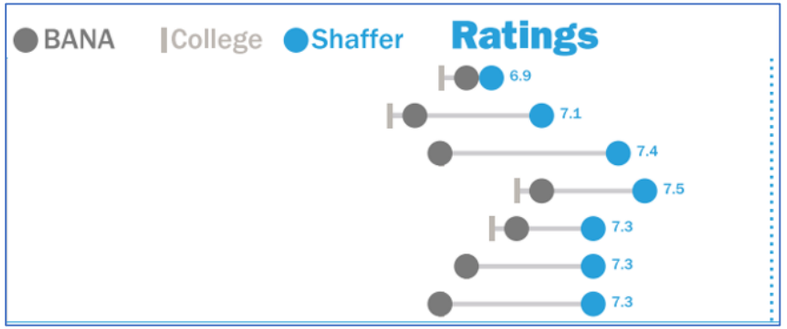
Fonte da Imagem: WEXLER, S., **SHAFFER**, J., & COTGREAVE, A. (2017)

Decisões de Design

- Barras de mesma largura



- Rotular somente uma das séries

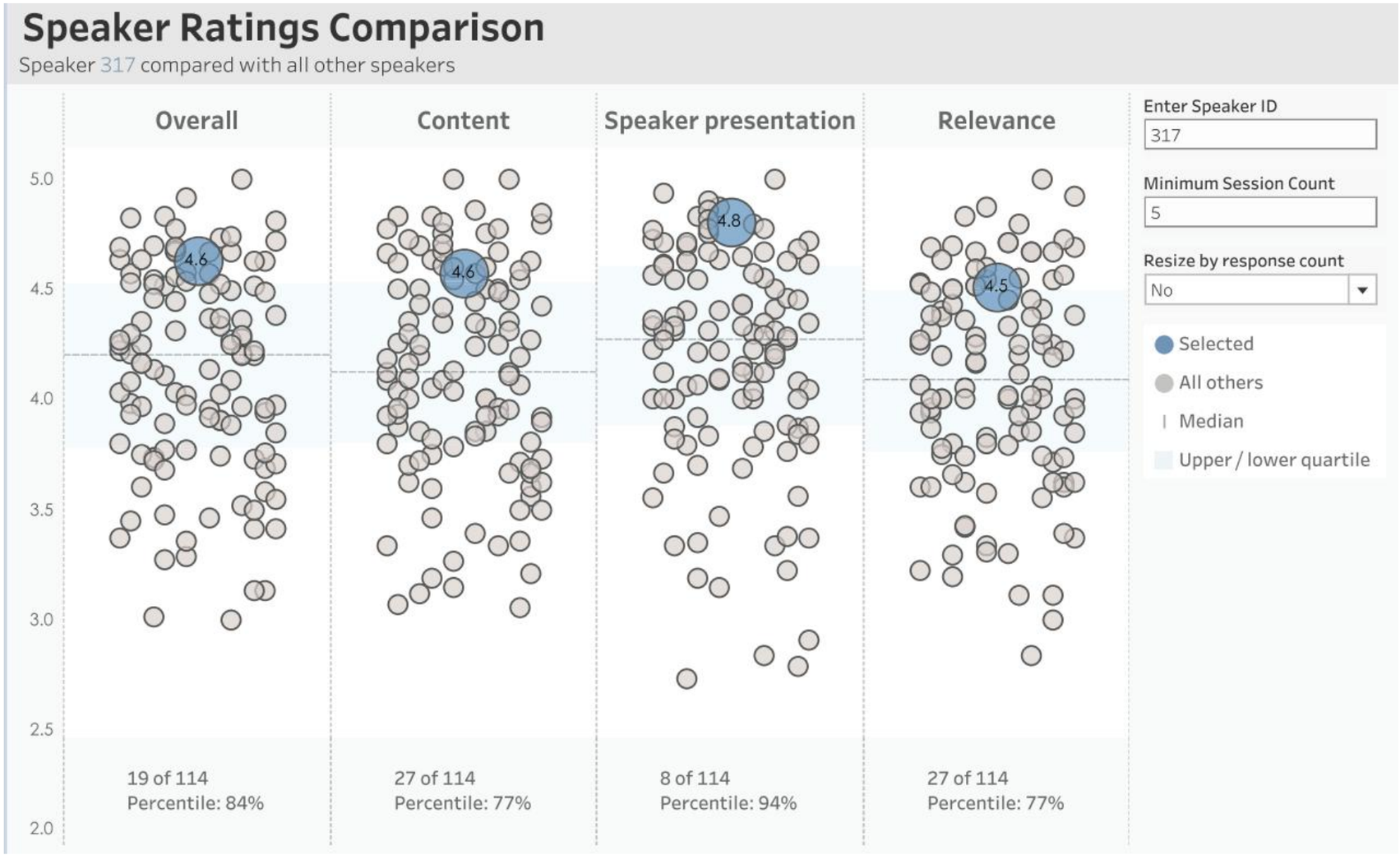


STEVE: Jeff's dot plot has become my go-to approach for comparing aggregated results from multiple sources (in this case an individual compared to a peer group compared to the college as a whole).

Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Dashboard de Oradores (jitterplot)

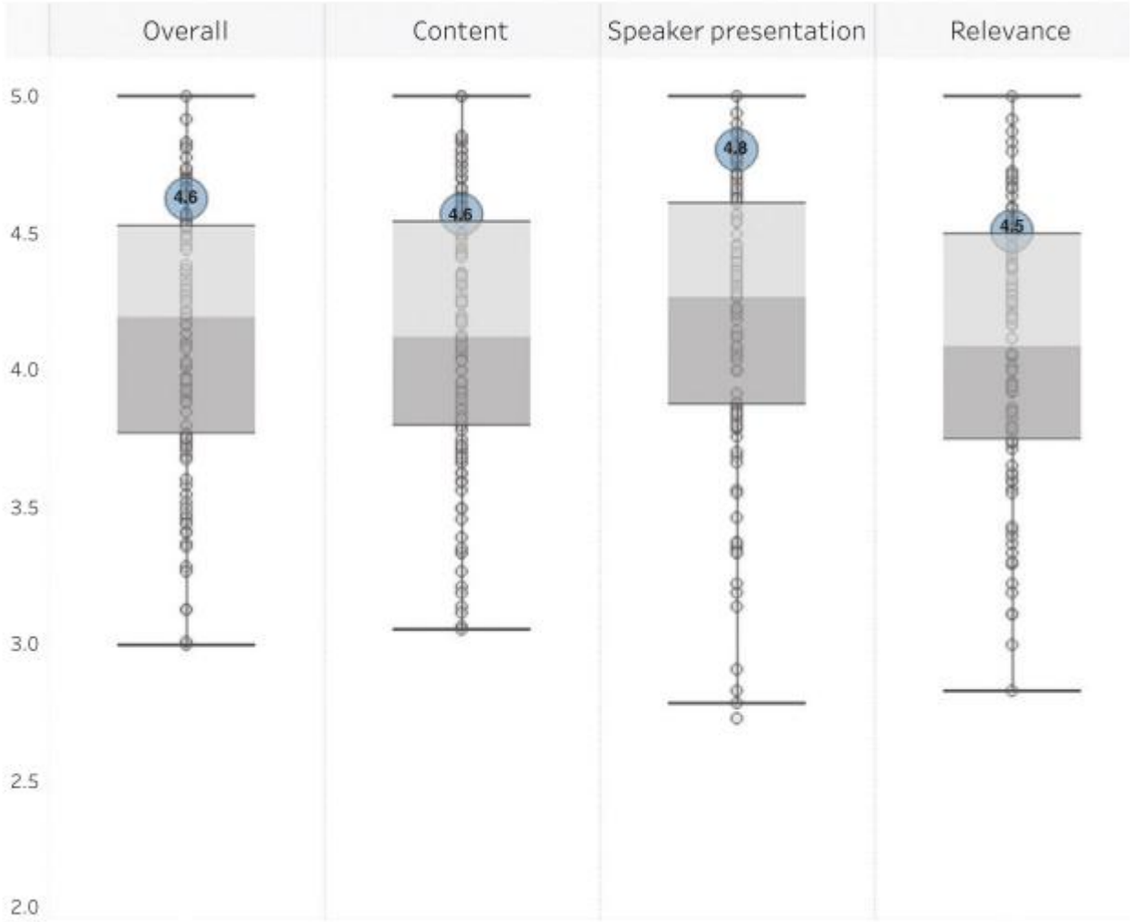
- A quantidade de pontos ajuda a contar uma história



Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Dashboard de Oradores

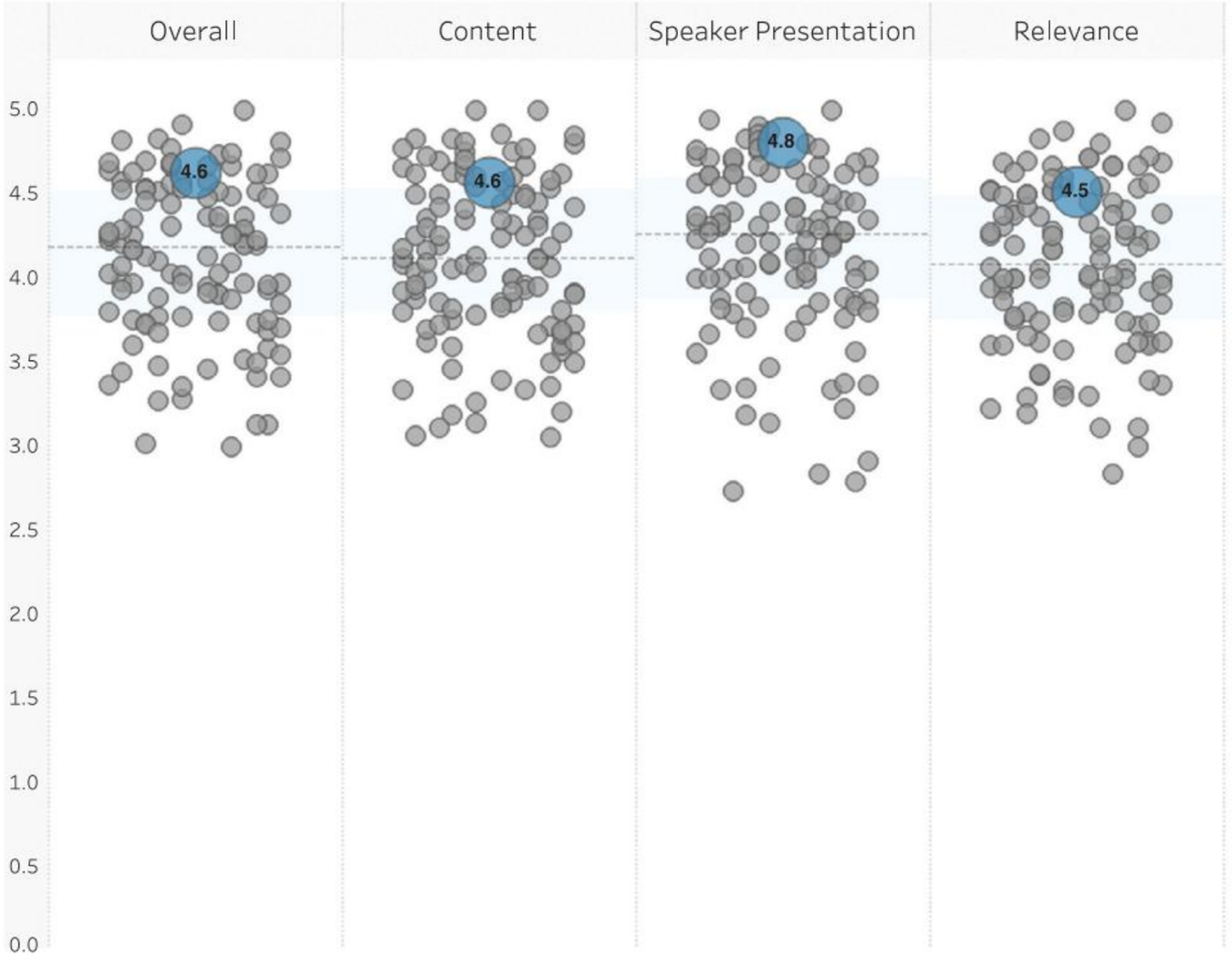
- Qual a desvantagem desta representação ?



Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

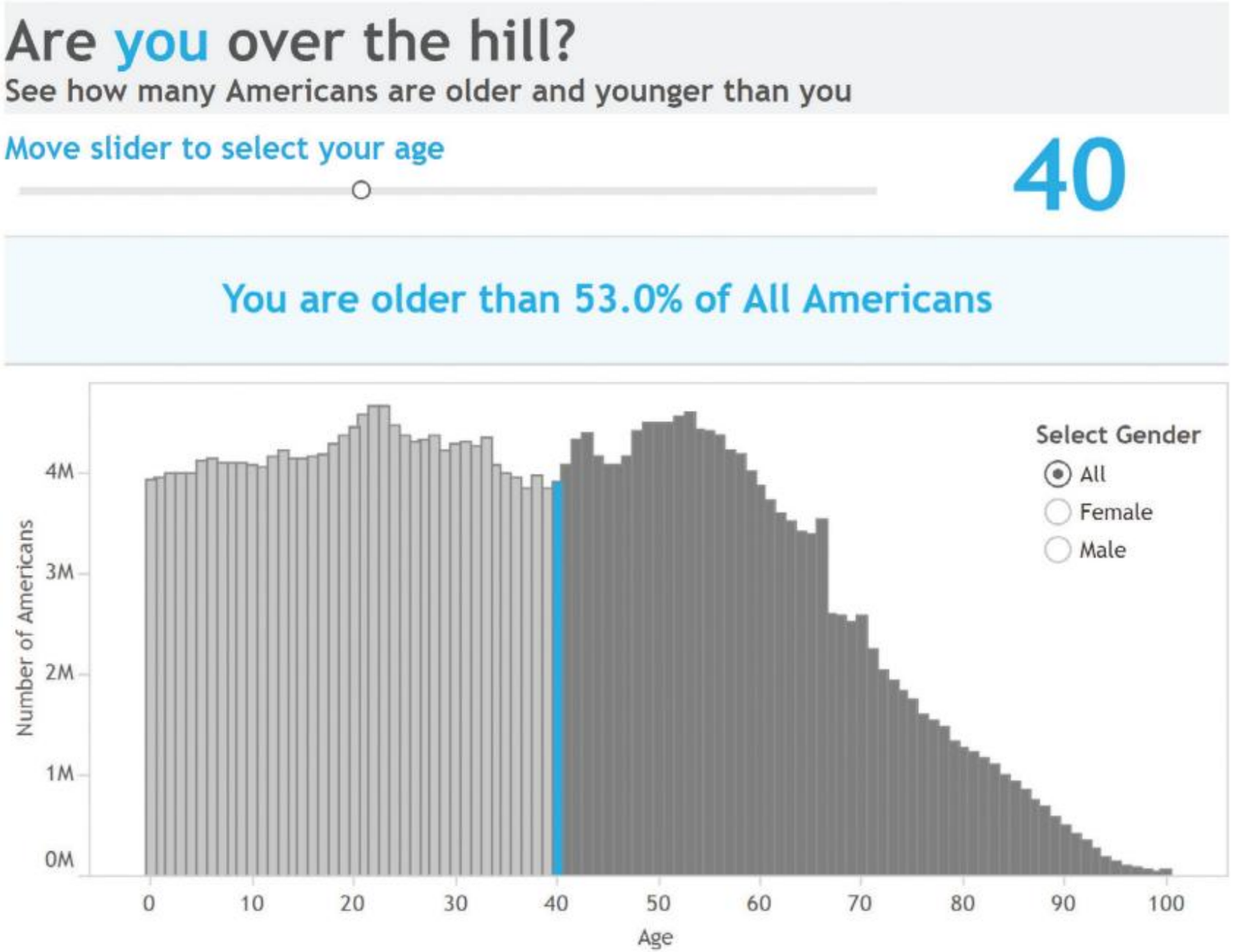
Dashboard de Oradores

- Usar ou não um eixo não começando no zero ?



Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

E se houver milhões de pontos ? Histograma



Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Dashboard de atendimento de saúde

- Visitas em casa e Admissões em UTIs
  - O gráfico de pontos resume um histórico clínico

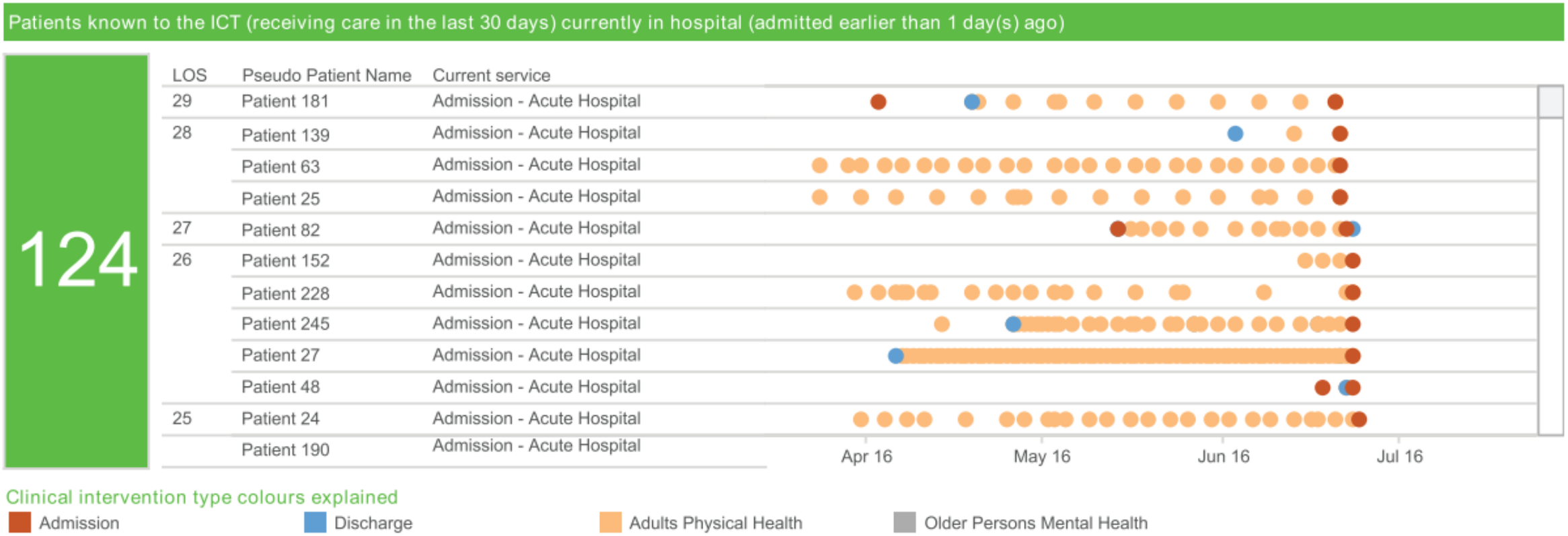


FIGURE 14.2 Patients admitted more than one day ago.

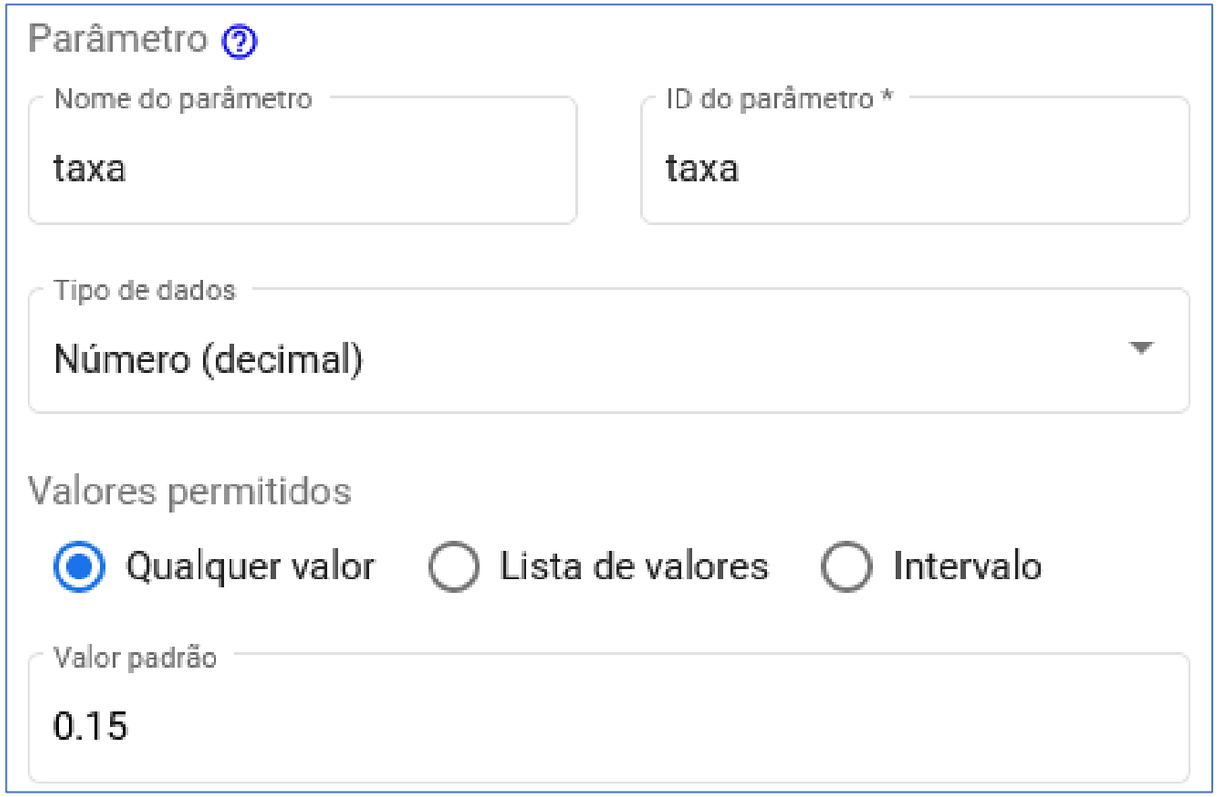
Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Google Looker Studio

https://support.google.com/datastudio/?hl=pt-BR

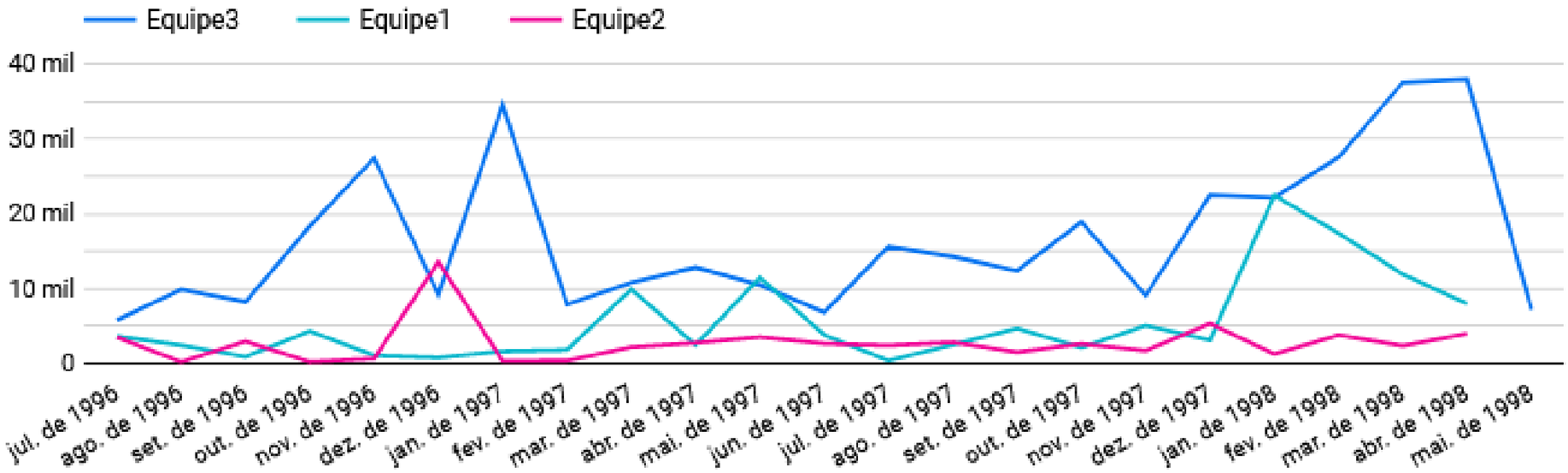
Parâmetro

- Útil para
  - Aplicar em campos calculados,
  - Enviados junto com uma query SQL (no BigQuery)
    - ✓ Por exemplo, quando se quer personalizar o data source a partir da interação com o usuário
- Podem receber dados
  - **De um valor padrão/estáticos**
    - ✓ Exemplo: população do Brasil
  - Do link para o relatório
  - De um campo de input
    - ✓ presente no relatório



Atividade 3.1 (5 min)

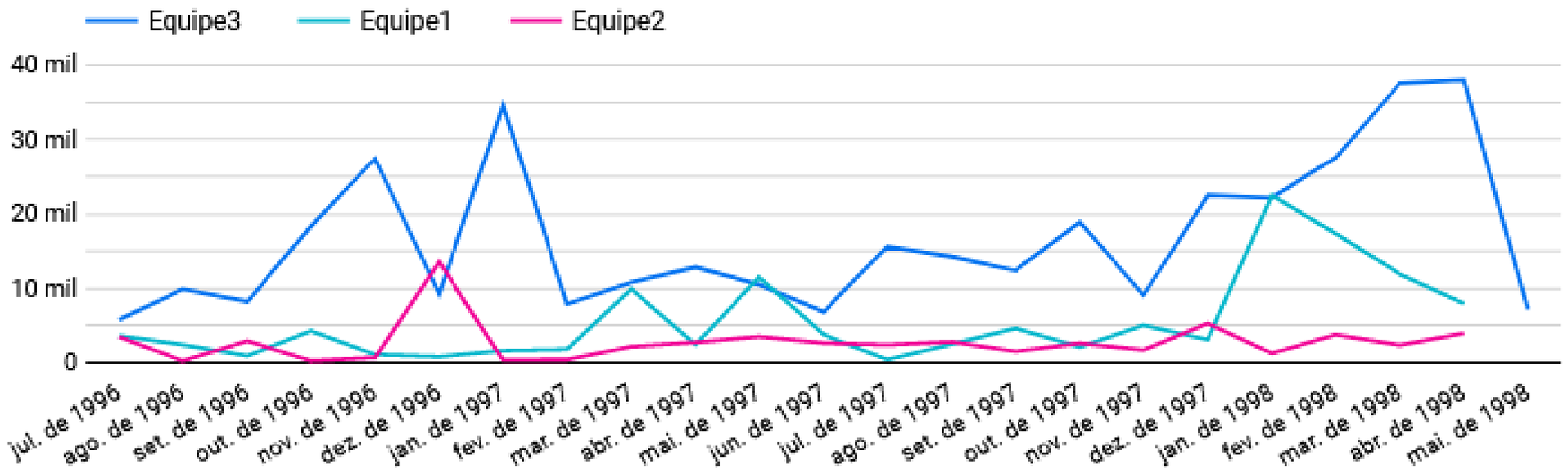
3.1) Criar um gráfico de linha do total de venda ($) por semana de cada equipe



- Resolução
  - Escolher a dimensão da data (OrderDate) e a métrica (subtotal)
  - Escolher a dimensão detalhada (equipe)

Uso propositalmente das cores

3.1) Criar um gráfico de linha do total de venda ($) por semana de cada equipe



- Demonstração

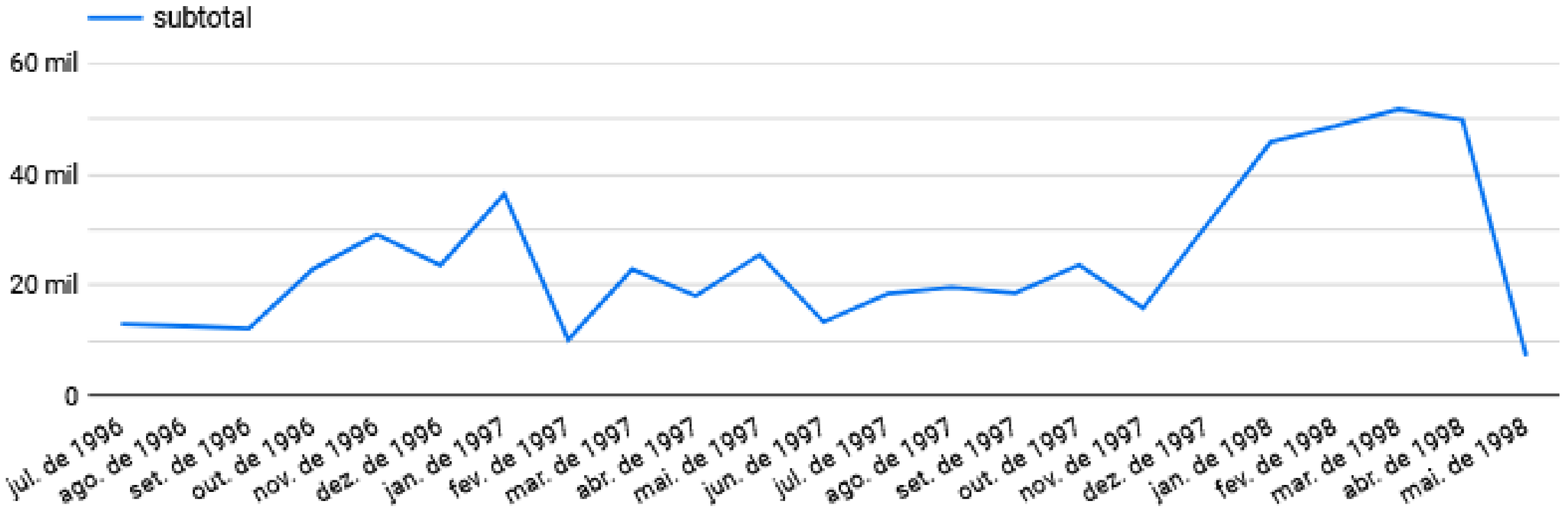
Removendo Clutter (Desordem)

	equipe	CategoryName	subtotal
1.	Equipe3	Beverages	212.600,25
2.	Equipe3	Seafood	97.734,02
3.	Equipe3	Produce	76.786,5
4.	Equipe1	Beverages	70.153,25
5.	Equipe1	Seafood	38.161,31
6.	Equipe2	Beverages	26.828,75
7.	Equipe2	Produce	21.016,75
8.	Equipe1	Produce	13.591,75
9.	Equipe2	Seafood	13.164,2

1 - 9 / 9

- Demonstração
  - Remover casas decimais
  - Resumir os números

Removendo Clutter (Desordem)



- Demonstração
  - Remover as linhas horizontais e simplificar as datas
- Função FORMAT_DATETIME

Exercício 3.1 (As atividades da próxima aula serão baseadas nesse dashboard)

- Crie um Dashboard no Google Looker Studio,
  - Contendo informações sobre a população e o PIB dos municípios
    - ✓ Uma **tabela** com informações básicas do ano de 2018
      - UF, Nome do Município, População, PIB e PIB Percapita
    - ✓ Um **gráfico** de linha com o PIB percapita sparado para cada um dos estados de SP, RS, CE e AM, dos anos de 2002 a 2018.
    - ✓ Uma **tabela** de heatmap com o PIB percapita dos estados (eixo vertical) ao longo dos 10 últimos anos (eixo horizontal)
  - **Aplique** os conceitos de Storytelling estudados nas aulas
- Use os dados do projeto Base dos Dados, hospedados no BigQuery
  - como fontes de informação para o seu modelo de dados
    - ✓ **Explique** textualmente a sua Query no caderno Colab

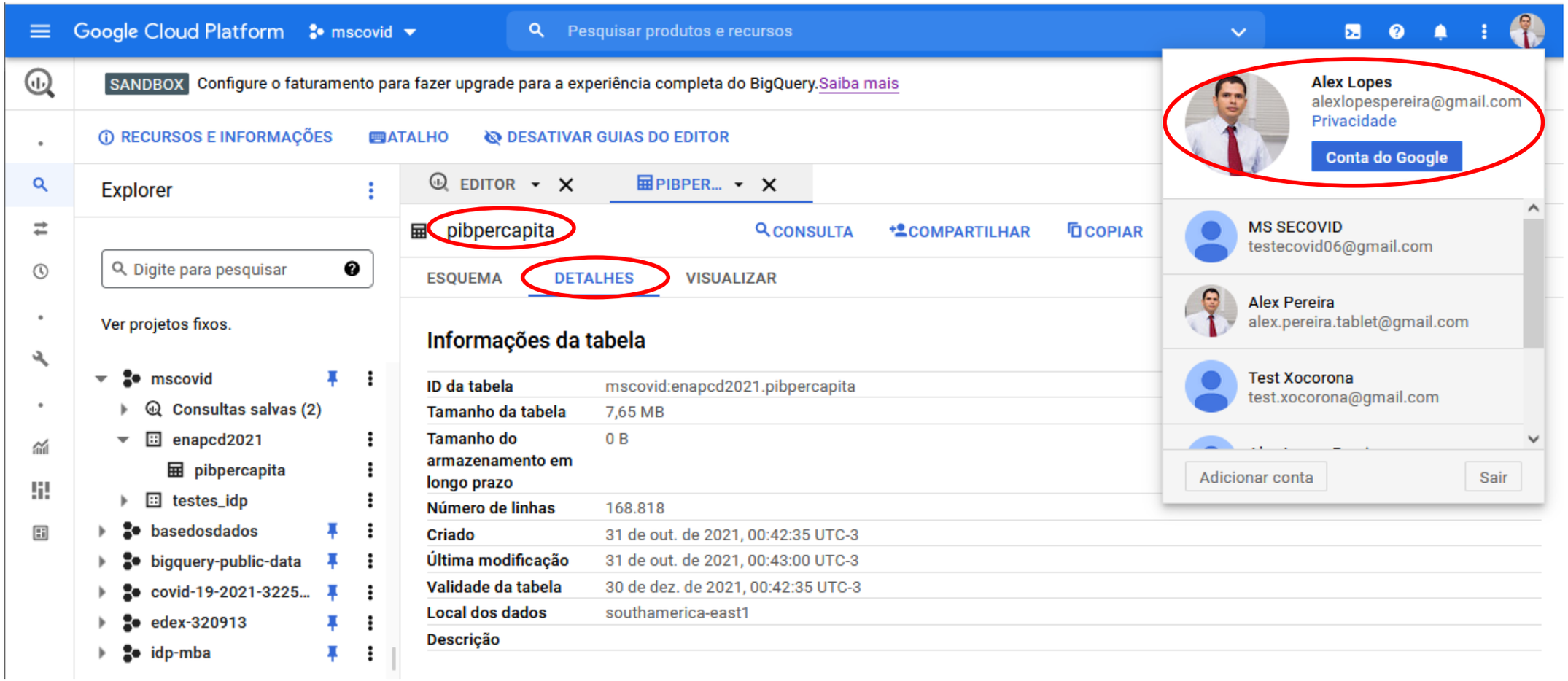
Exercício 3.1

- Use a metodologia apresentada na aula
  - **ETL com pandas** e criação de tabela no BigQuery, conforme modelo
- **Crie um data source** para contectar com sua tabela no BigQuery
- Submeta aqui as evidências do seu trabalho
  - Link **público** do dashboard, print do BigQuery conforme o modelo do próximo slide, e o link **público** para o seu caderno Colab (1/3 da nota).
    - ✓ **Não** compartilhe com meu email
      - Demonstração
  - Link público para o seu notebook no colab ou github (1/3 da nota)
    - ✓ Teste numa aba anônima
  - Link público para um vídeo seu explicando como resolveu o problema (1/3 da nota)
    - ✓ Abordando conceitos da aula

Link privado = não entregue / atrasado

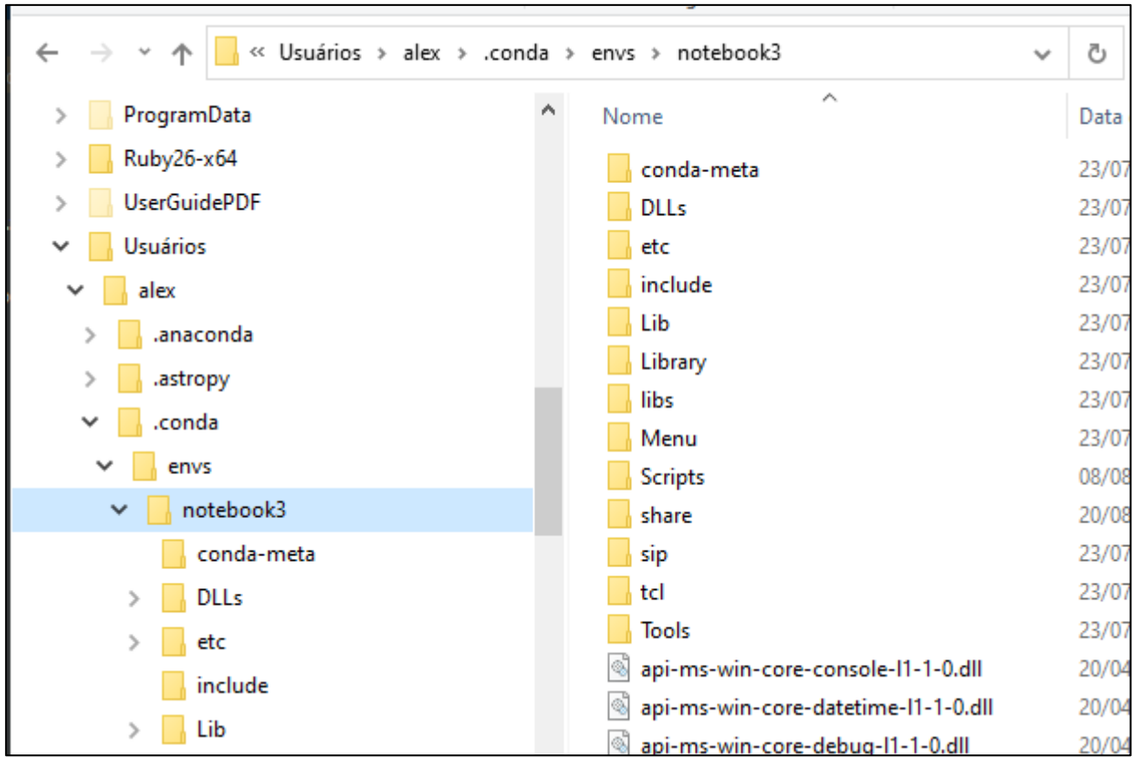
Exercício 3.1

- Modelo de screenshot (print) do BigQuery
  - Apresente uma imagem contendo os detalhes destacados



Ambiente Virtual (Virtual Environment)

- Ambiente virtual
  - é um ambiente isolado de pacotes python
    - ✓ Elimina problema de diferentes projetos precisarem de várias versões do mesmo pacote
    - ✓ Minimiza risco de provocar instabilidade nas ferramentas baseadas em python utilizadas em outros contextos
- Instalador de pacotes –
  - pip/conda (existem outros)
    - ✓ conda/pip install <PACOTE>
      - pip install pandas
      - conda install pandas



Configuração do Ambiente Virtual com pip

- Instalar o pacote virtualenv
  - `pip install virtualenv`
- Criar e entrar um diretório
  - `mkdir virtual_envs && cd virtual_envs`
- Criar um ambiente virtual chamado env
  - `python -m venv env`
- Executar o script activate pra habilitar o ambiente env
  - env\Scripts\activate
    - ✓ a partir deste momento qualquer pacote será instalado no ambiente virtual
      - enquanto o ambiente estiver ativado: (env) no início do prompt
  - env\Scripts\deactivate.bat
- Instalar um pacote de teste (`pip install etlelk`)
- Crie um ambiente virtual para cada projeto

Introdução ao VS Code / Windsurf

- Download e instalação do Python
  - Demonstração
- Criar um novo arquivo python
  - Demonstração
- Criar um ambiente virtual
  - Demonstração
- Clonar um repositório público (seu ou de outra pessoa)
- Depurar (debugar) o seu código
  - Criar um breakpoint e executar o arquivo em modo debug
    - ✓ Inspecionar as variáveis
    - ✓ adicionar uma expressão aos Watches
    - ✓ Avancar uma linha com o botão step over
    - ✓ Aprender a usar os outros botões da barra de debug
- Instalação de pacotes no terminal
  - pip install selenium
  - pip install -r requirements

Teste a utilização do Selenium – Pré-requisito para a próxima aula

- Selenium é uma biblioteca de web scraping
  - Se você instalou os pacotes usando o comando pip (slide anterior)
    - ✓ O Selenium já está disponível para uso no seu projeto
- Ateste o correto funcionamento do Selenium com Debug (breakpoint)
  - Executando este script