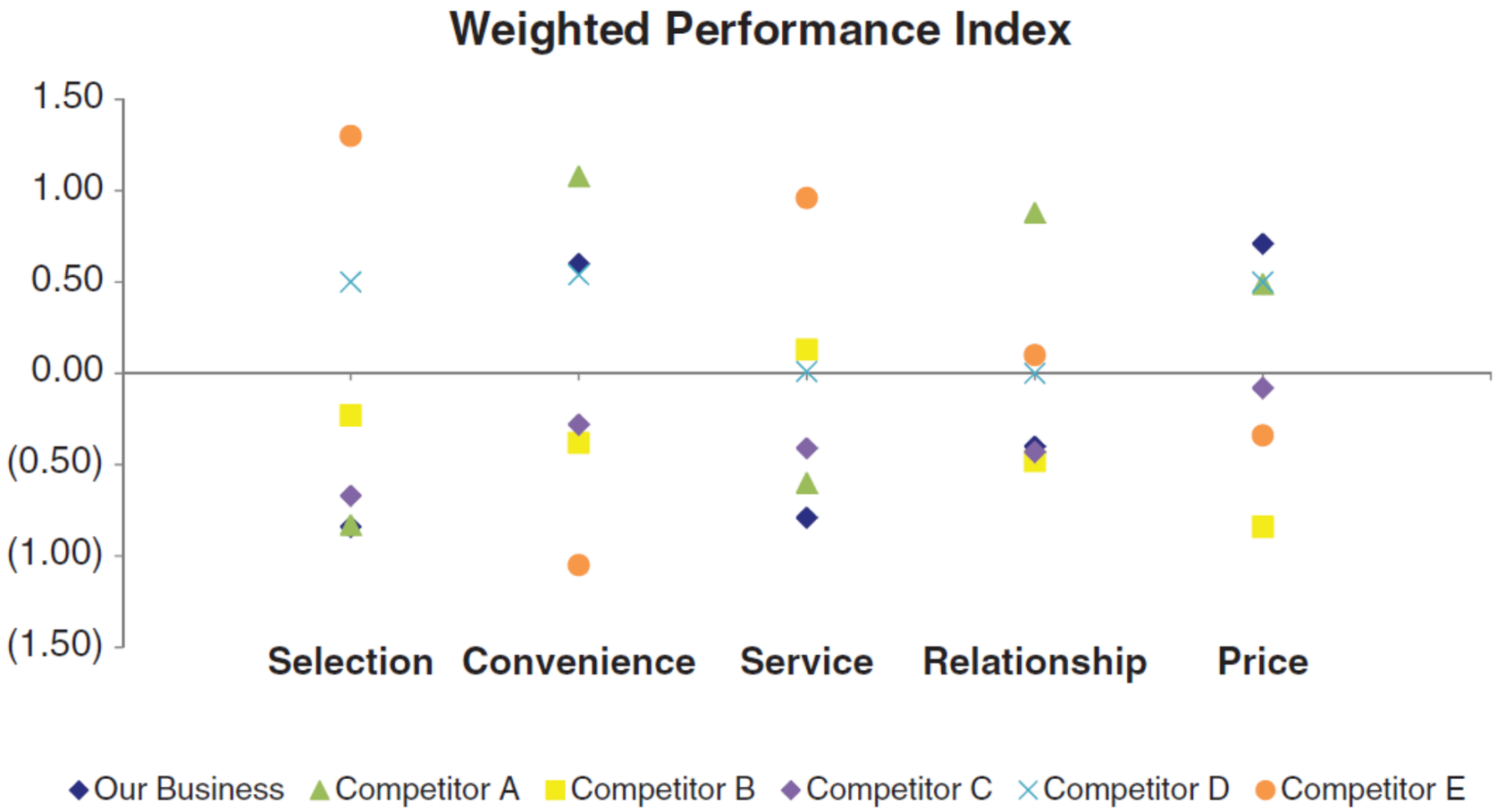
Introdução a Ciência de Dado

Professor: Alex Pereira

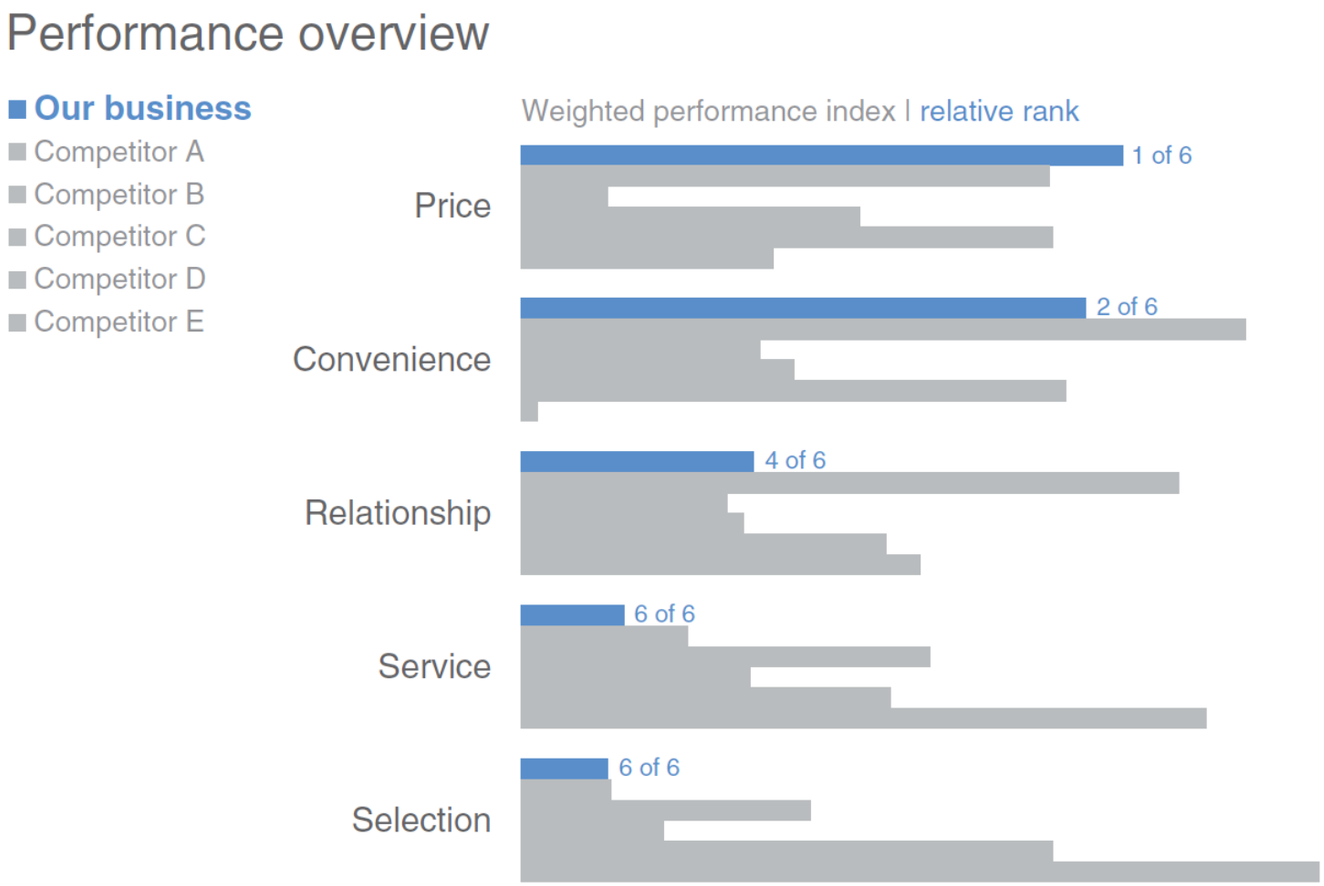
Uso não estratégico/proposital do contraste/cores

- O contraste ajuda a audiência a focar no que é importante



Fonte: KNAFLIC, C. N. (2018)

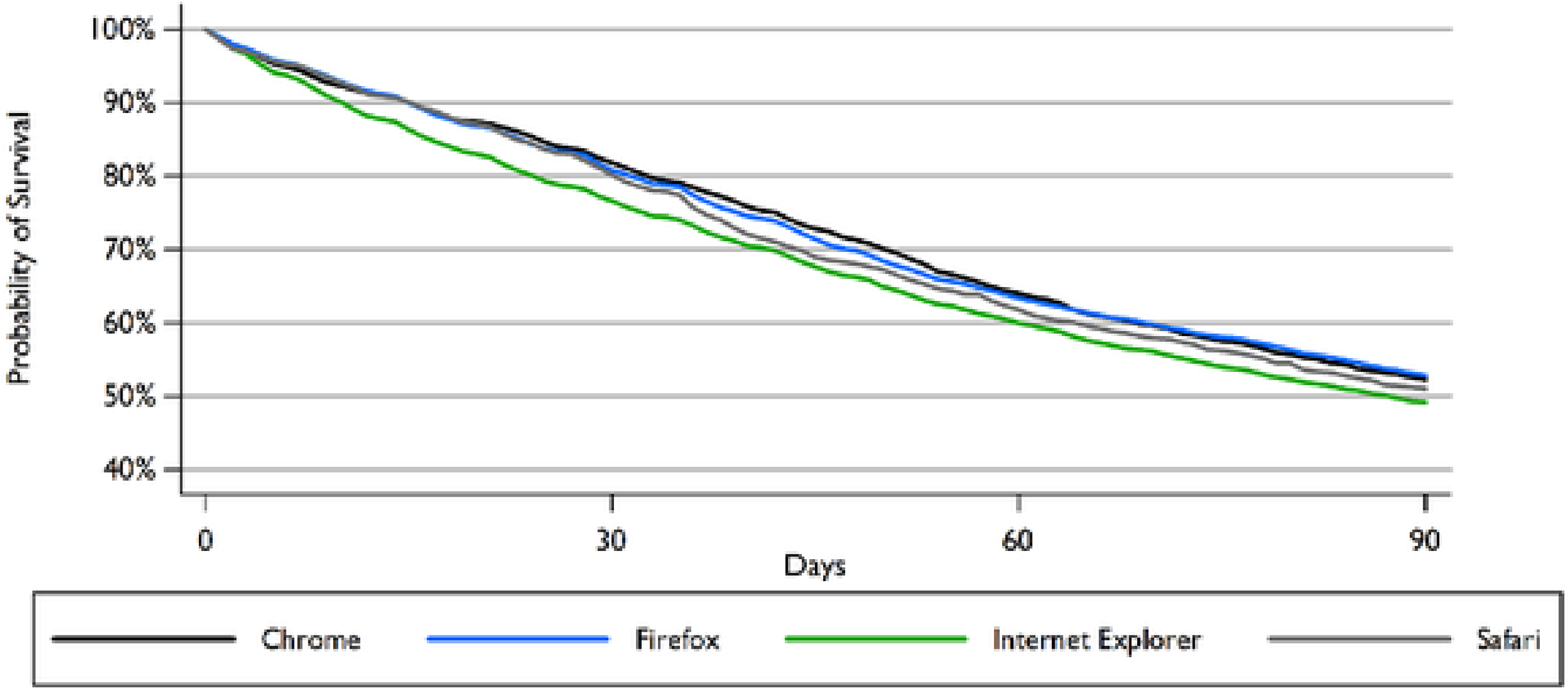
Uso estratégico/proposital do contraste/cores



- Assimilar a mesma mensagem num gráfico colorido sem propósito
  - Demanda esforço cognitivo para ignorar os estímulos visuais.

Fonte: KNAFLIC, C. N. (2018)

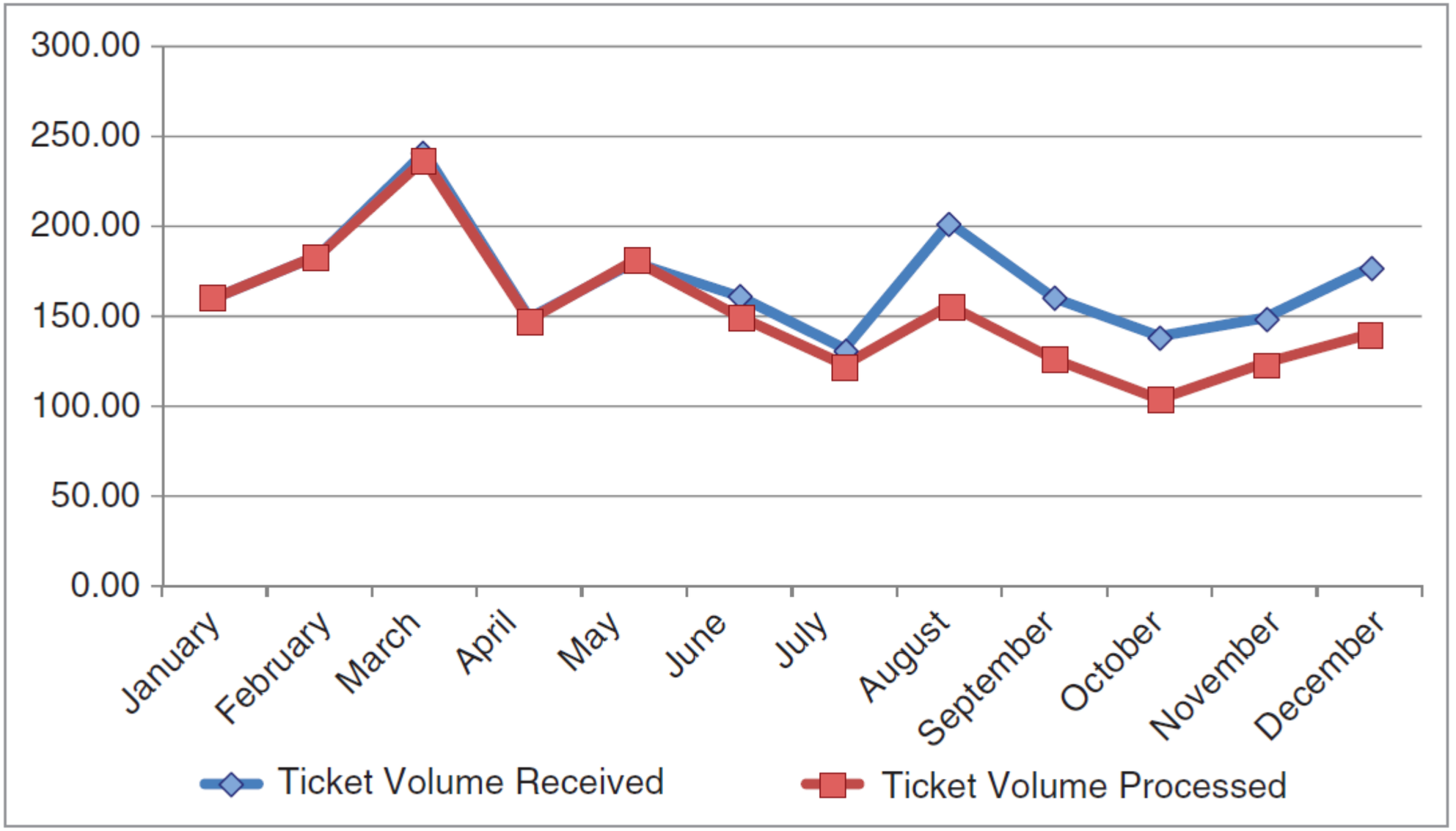
Probabilidade de Permanecer no Emprego VS seu Navegador (Questione o valor padrão)



Fonte

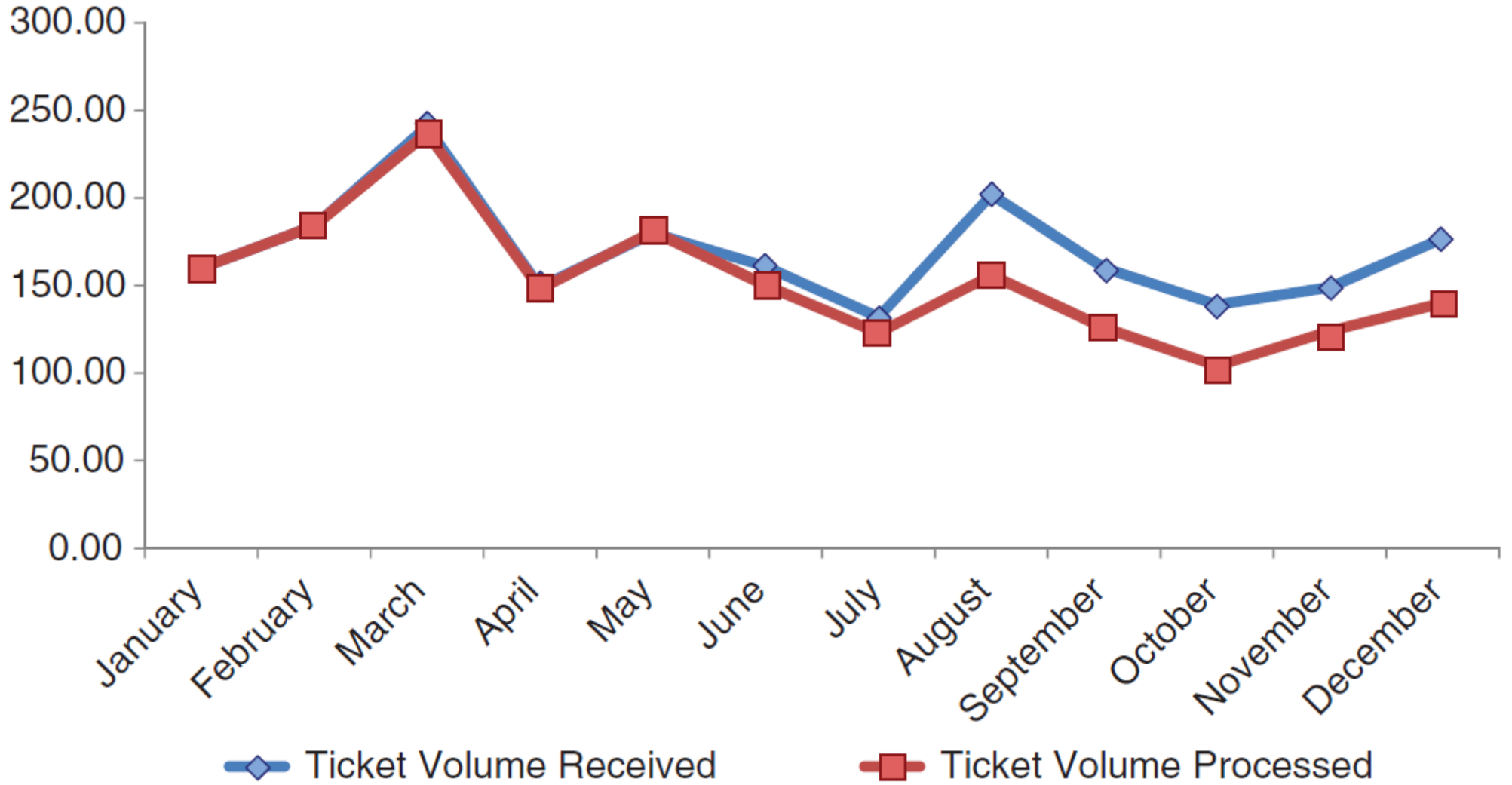
Artigo da Harvard Business Review sobre default

Removendo Clutter (desordem)



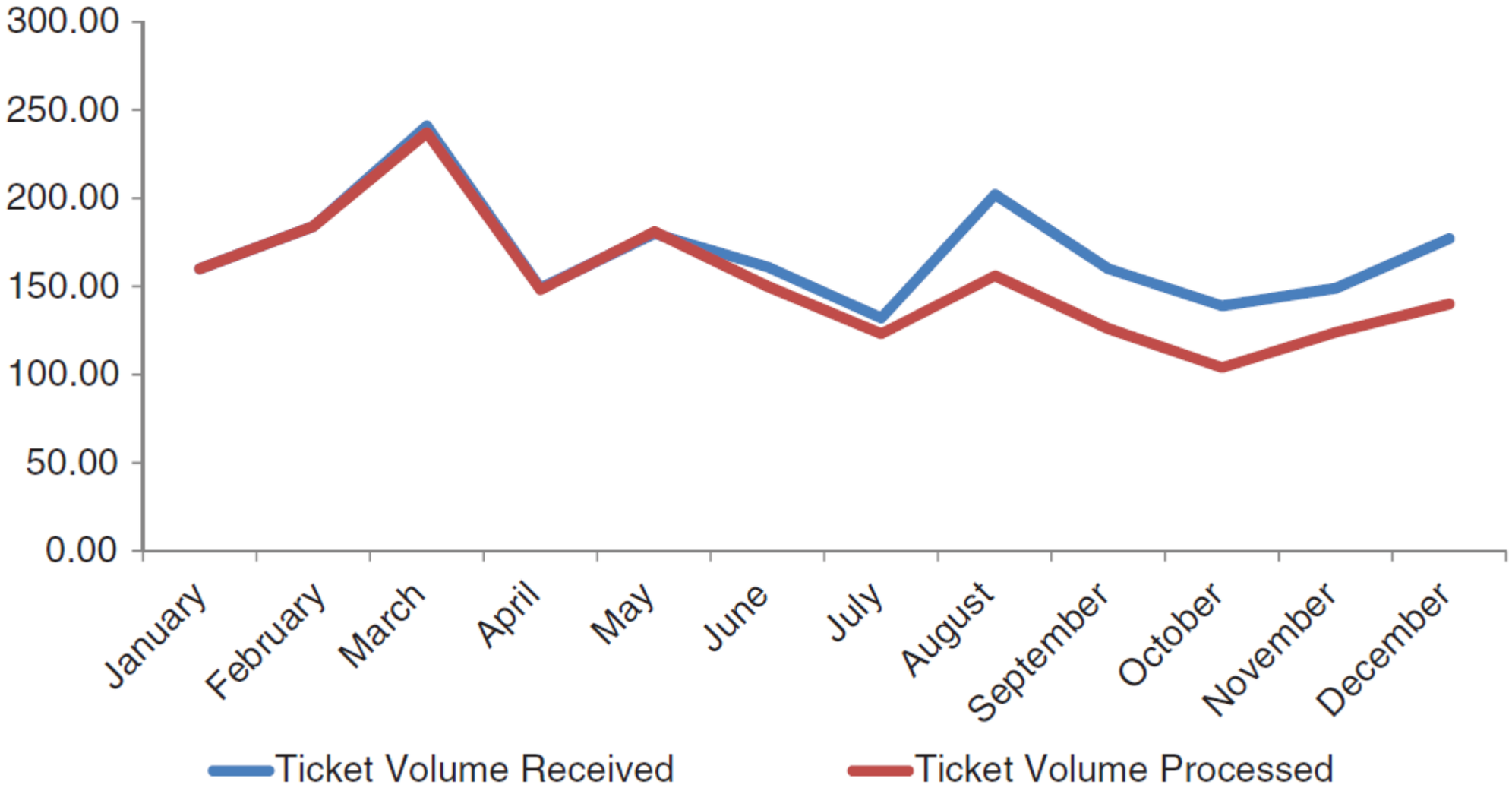
Fonte: KNAFLIC, C. N. (2018)

Removendo Clutter (desordem)



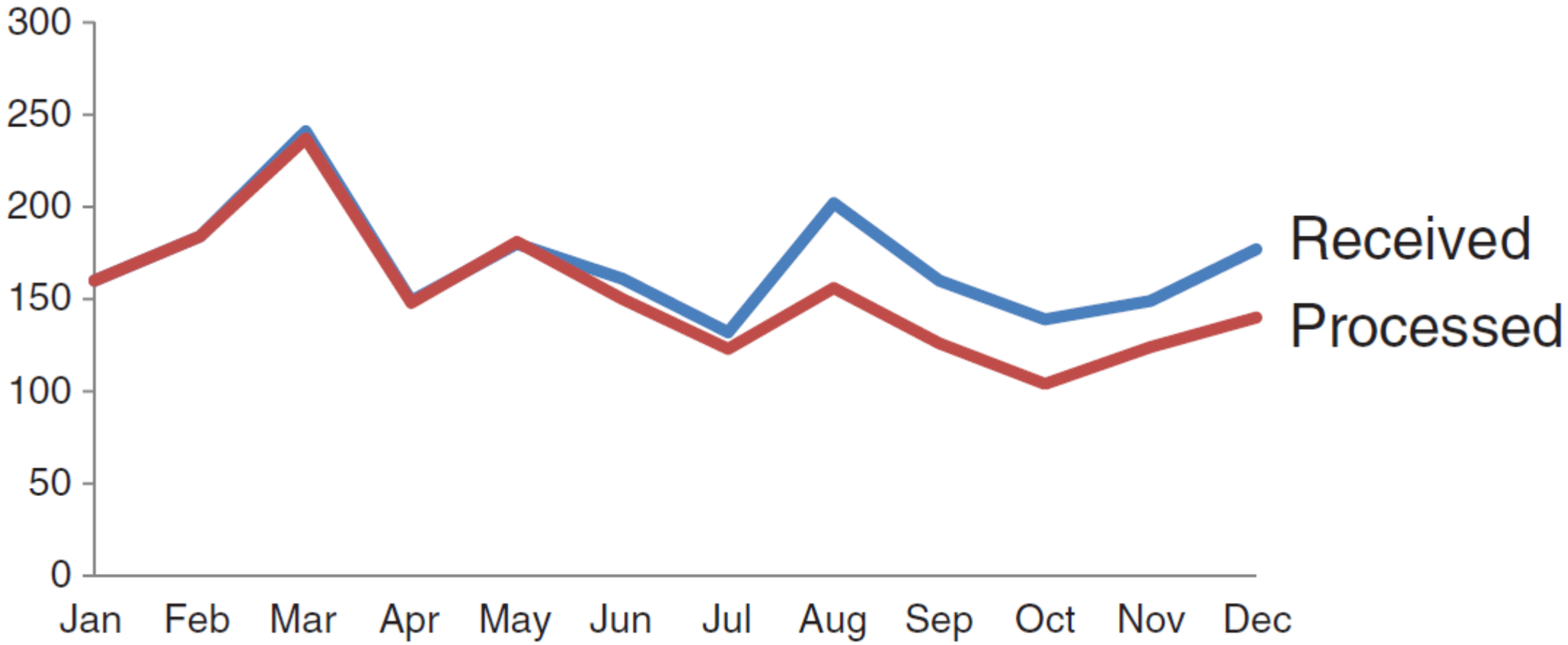
Fonte: KNAFLIC, C. N. (2018)

Removendo Clutter (desordem)



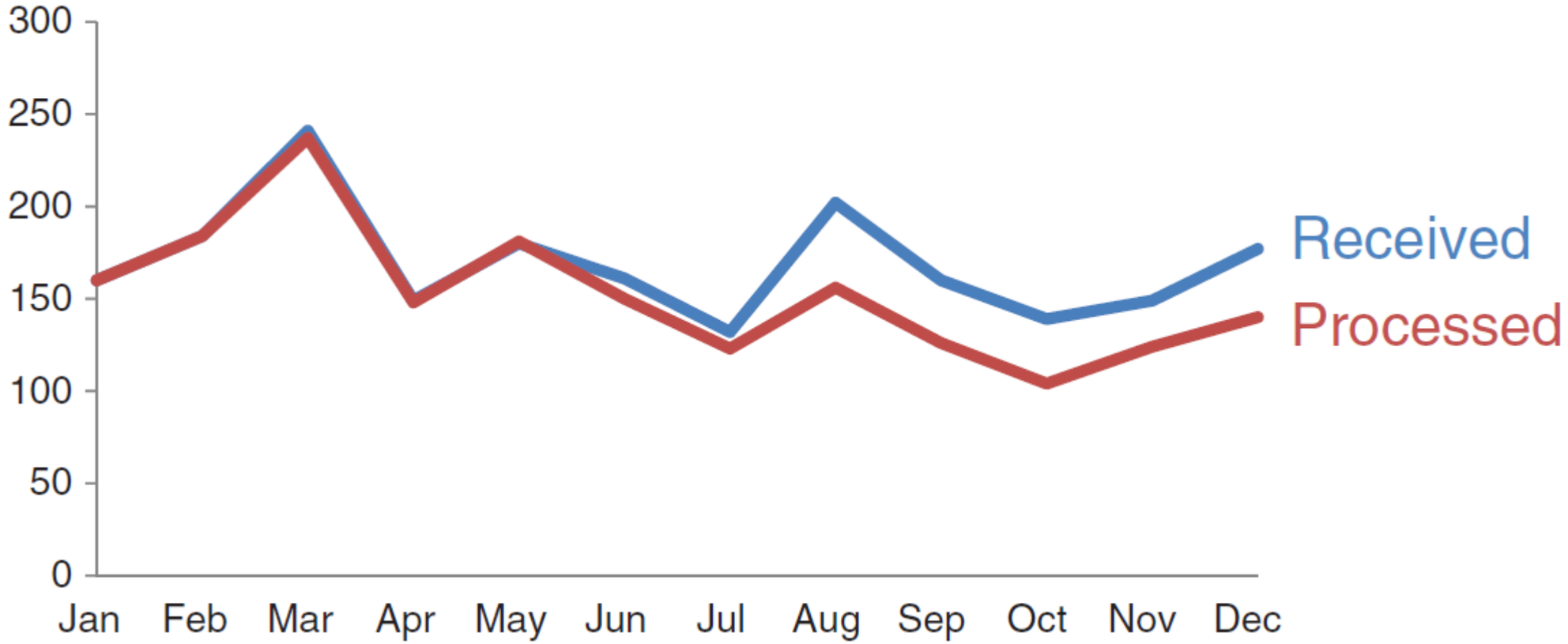
Fonte: KNAFLIC, C. N. (2018)

Removendo Clutter (desordem)



Fonte: KNAFLIC, C. N. (2018)

Removendo Clutter (desordem)



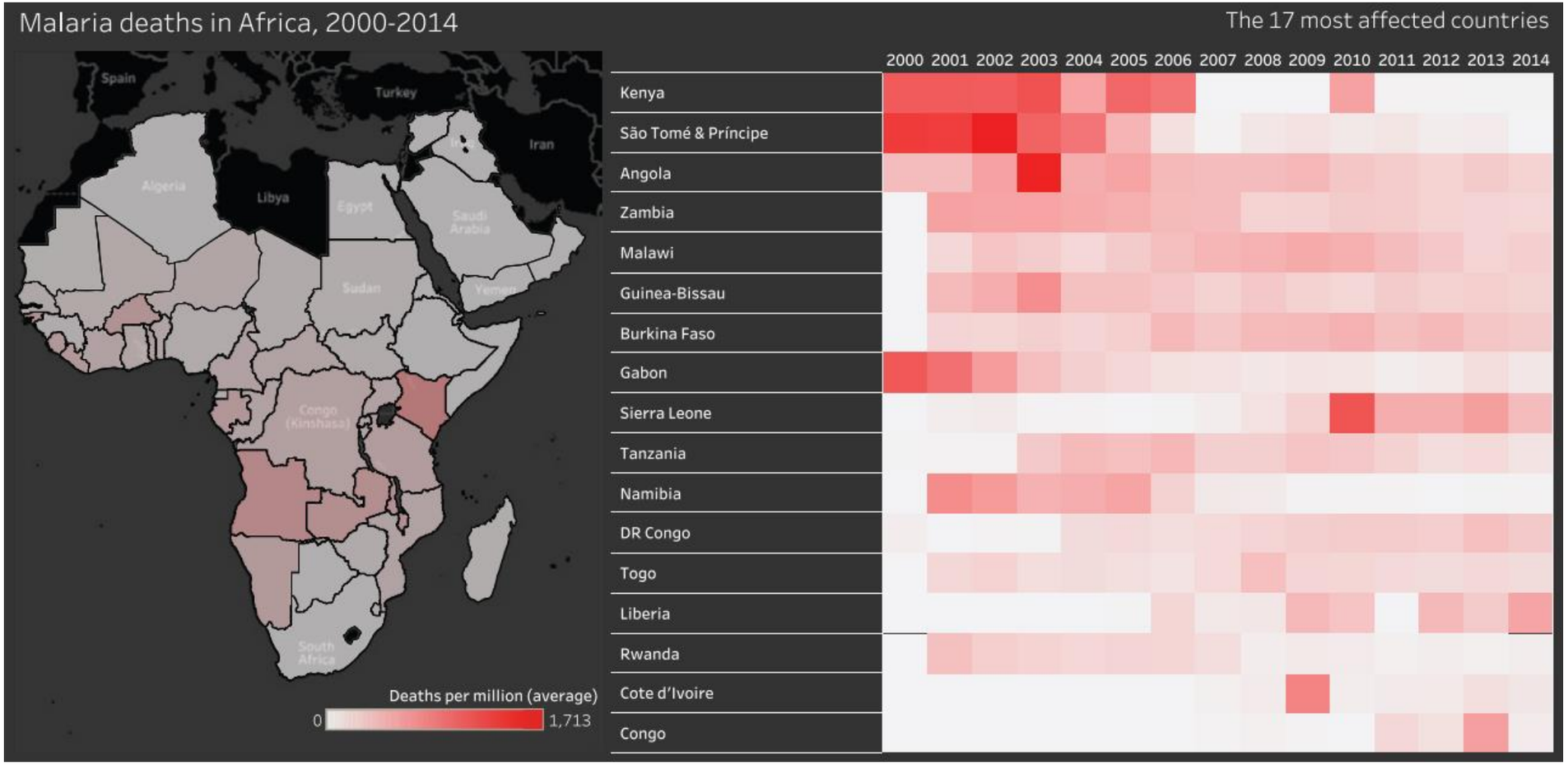
Fonte: KNAFLIC, C. N. (2018)

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.

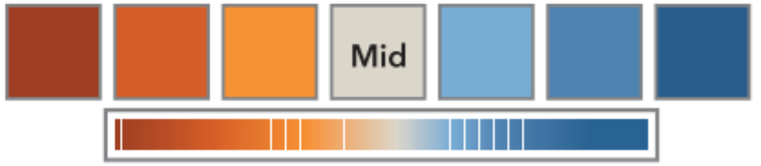
Data	Data Type	Encoding	Note
Country	Categorical	Position	The map shows the position of each country. In the highlight table, each country has its own row.
Deaths per million	Quantitative	Color	The map and table use the same color legend to show deaths per million people.
Year	Ordinal	Position	Each year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

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

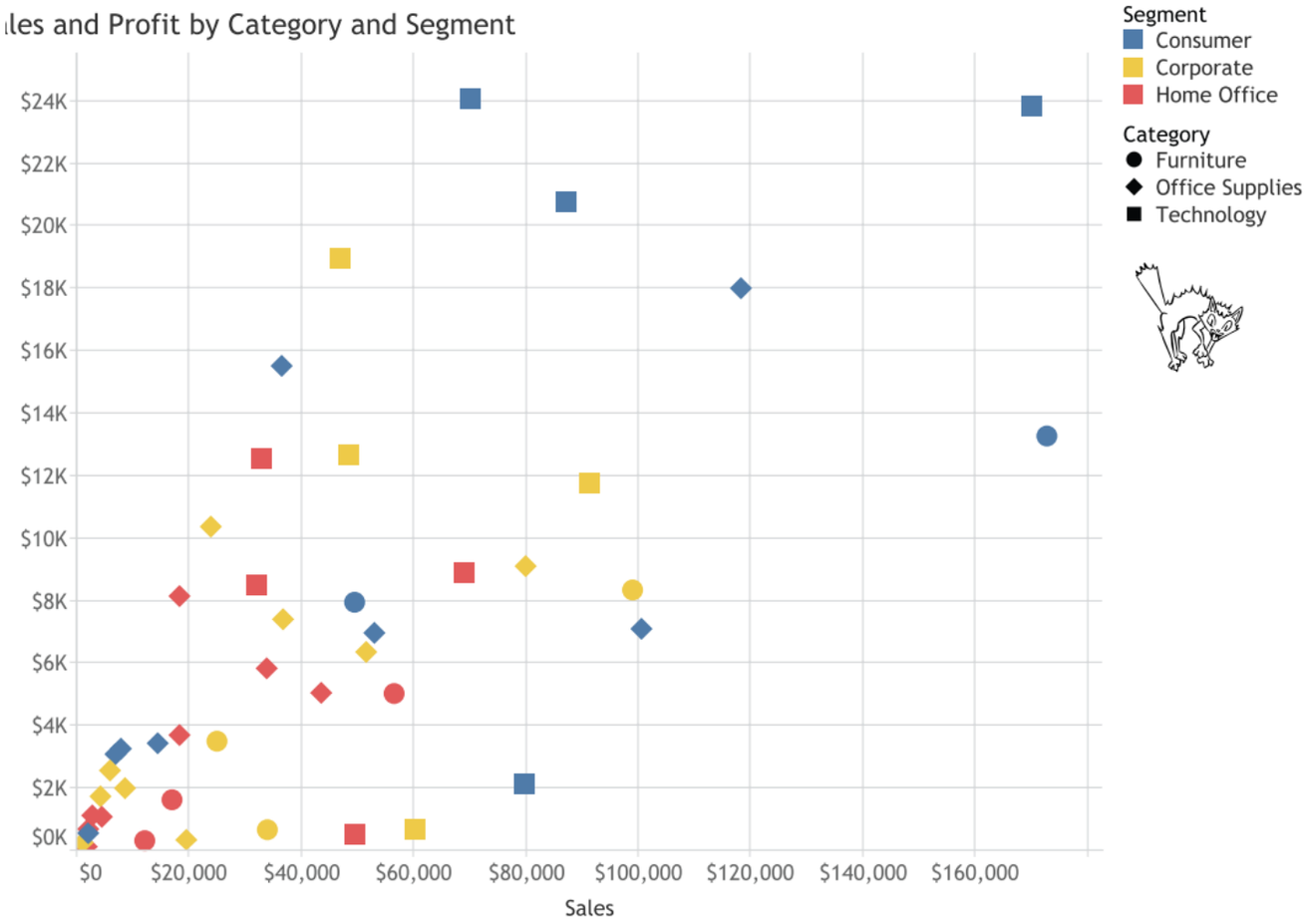


FIGURE 1.42 Scatterplot using shape and color. Which category has the highest profits?

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).

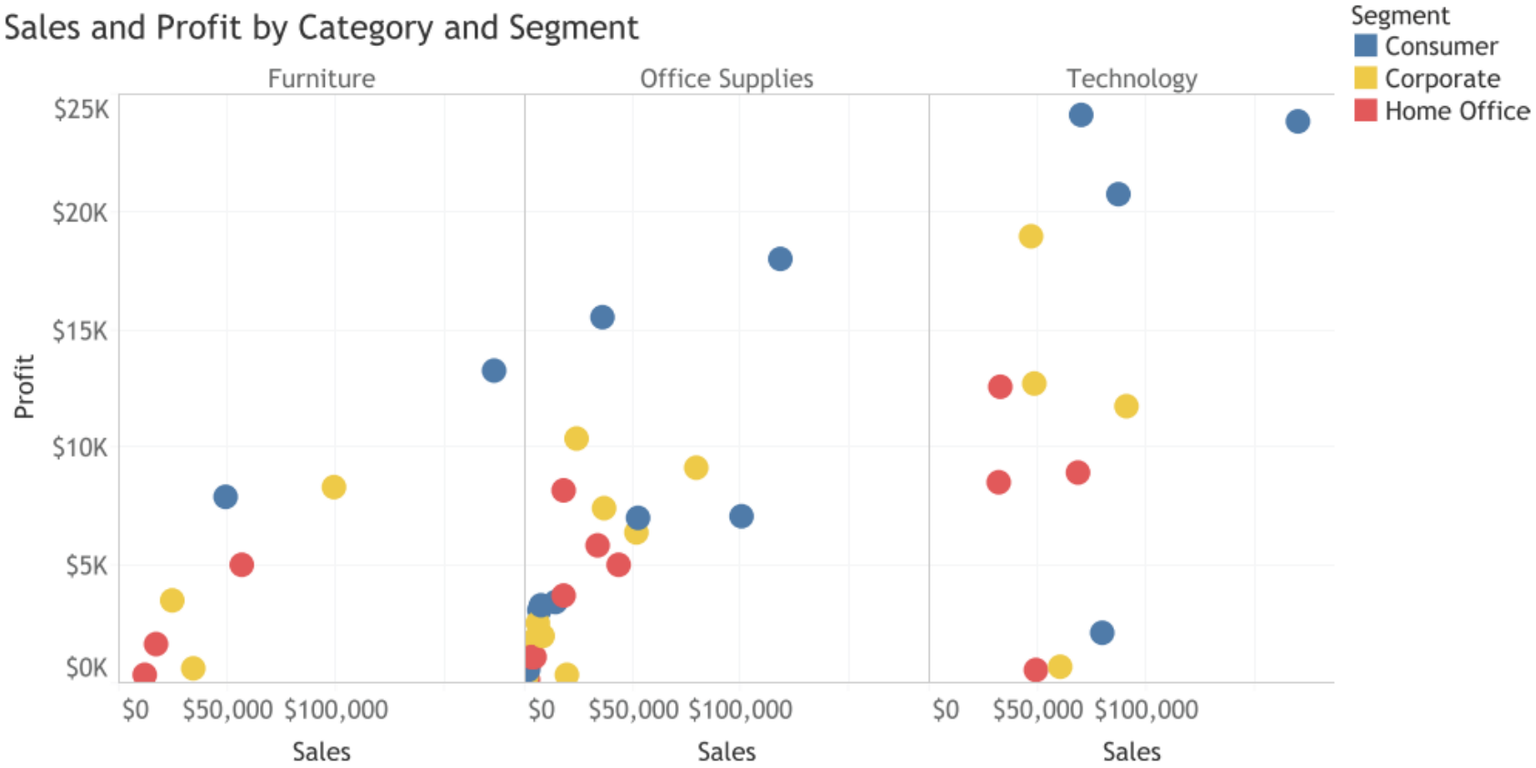


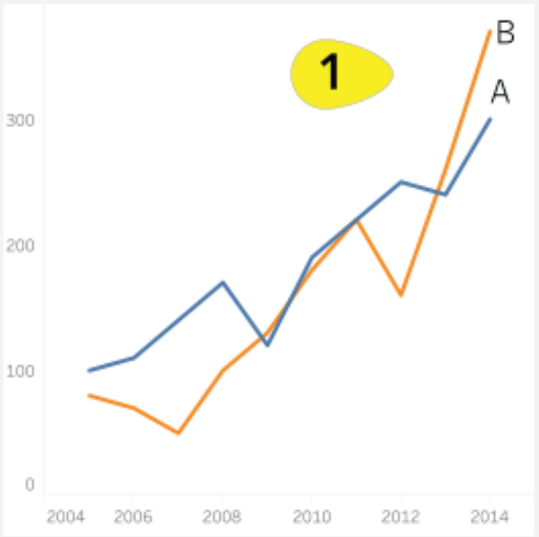
FIGURE 1.43 Sales and profit with one column for each category.

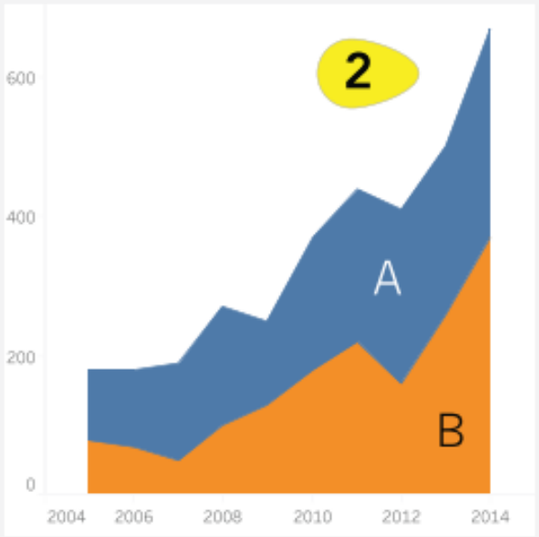
Fonte da Imagem: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

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





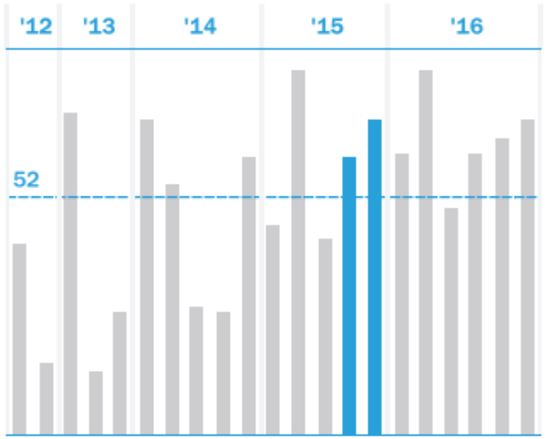
- Resposta pra quase tudo: depende da audiência
  - Quem são e seus objetivos.

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

Exemplo de Dashboard de uma Universidade

Course Metrics

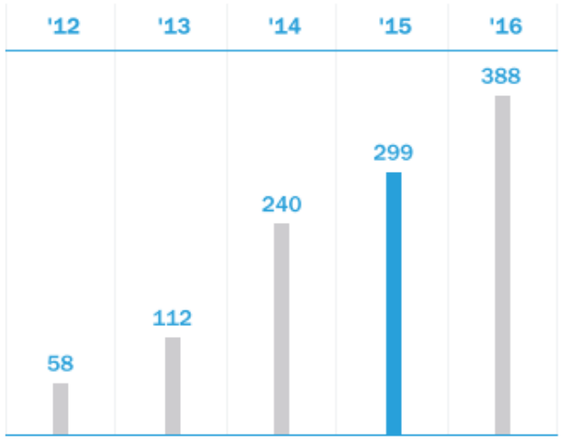
Students



1097

Total students in five years

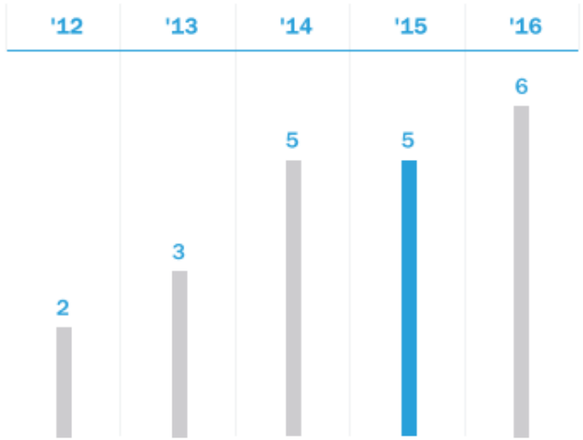
Enrollments



687

Total students in 2015-2016

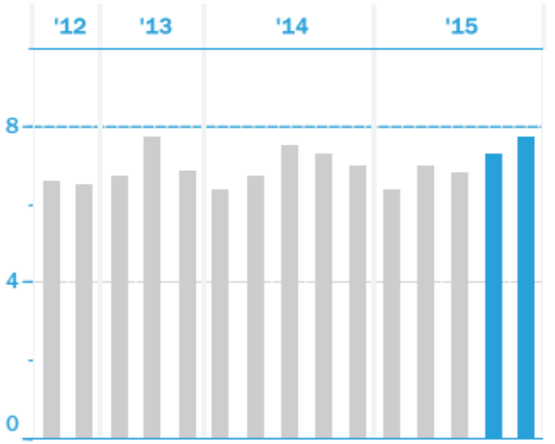
Classes



21

Total classes in five years

Ratings

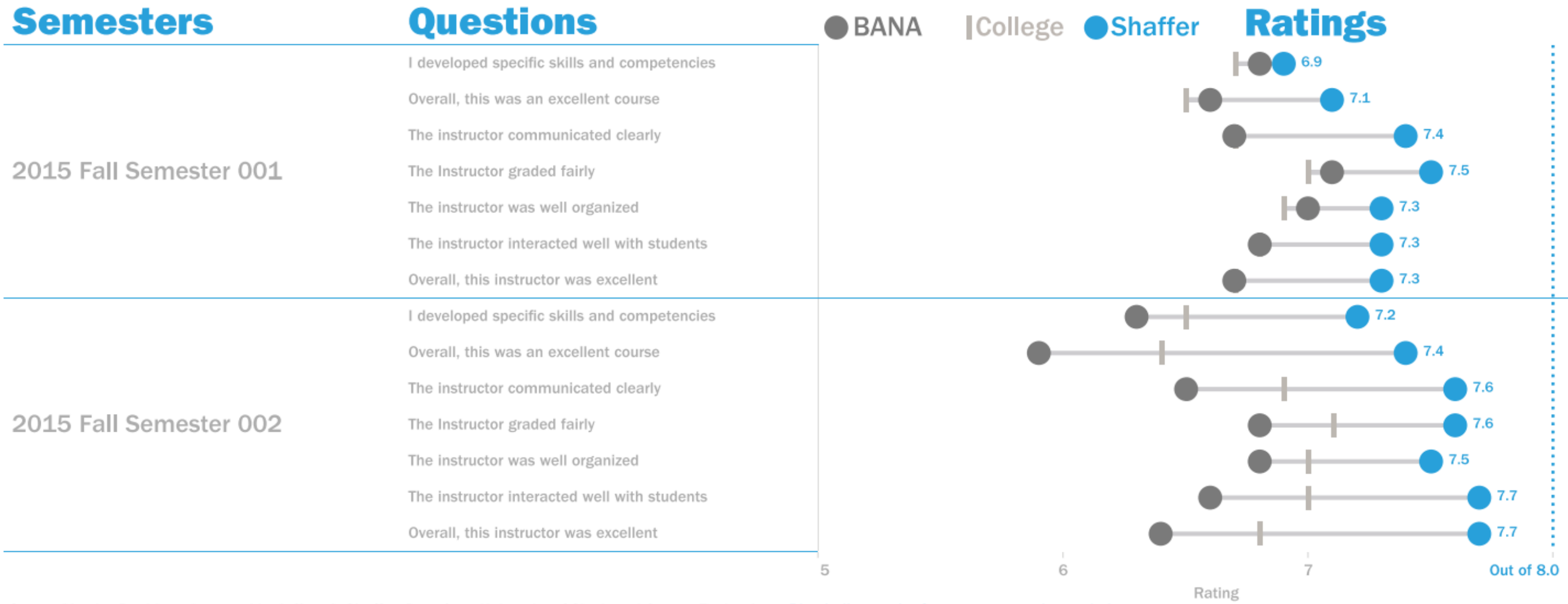


7.7 of 8

Most Recent Instructor Rating (out of 8.0)

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

Exemplo de Dashboard de uma Universidade

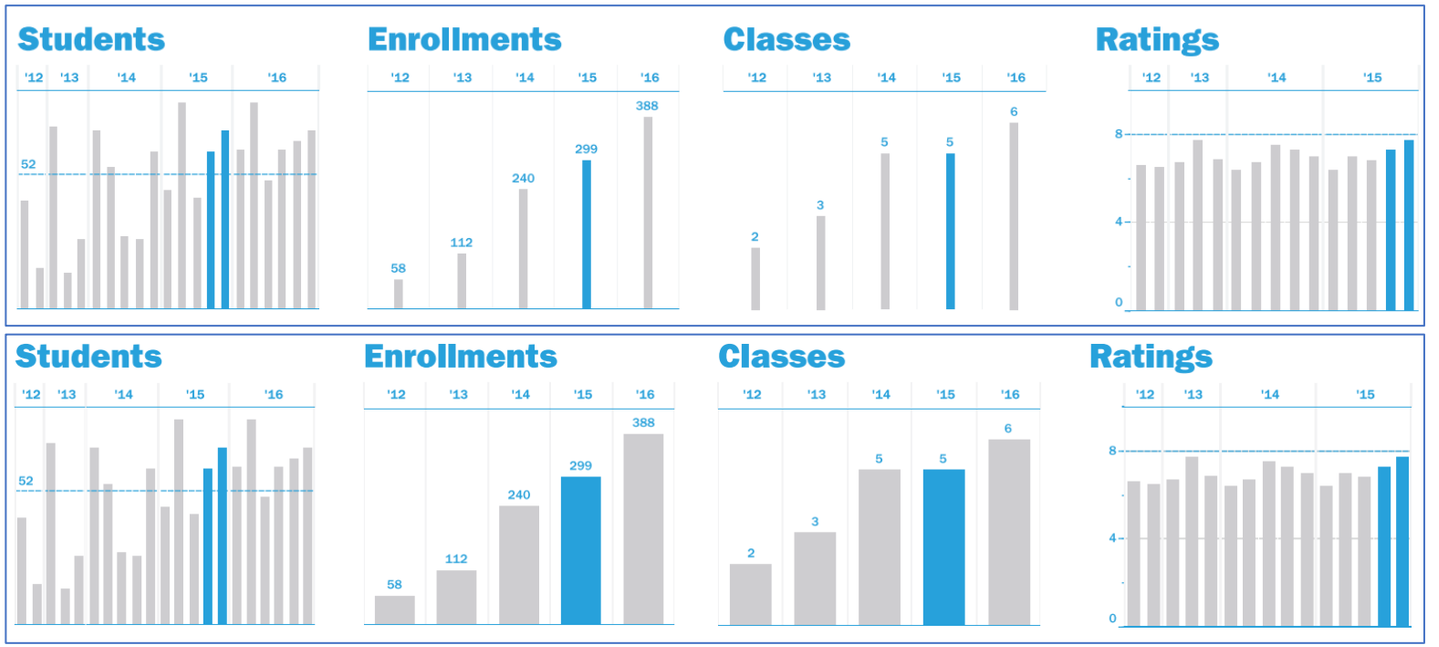


* BANA = Business Analytics; Shaffer = Professor do Curso de Visualização de Dados

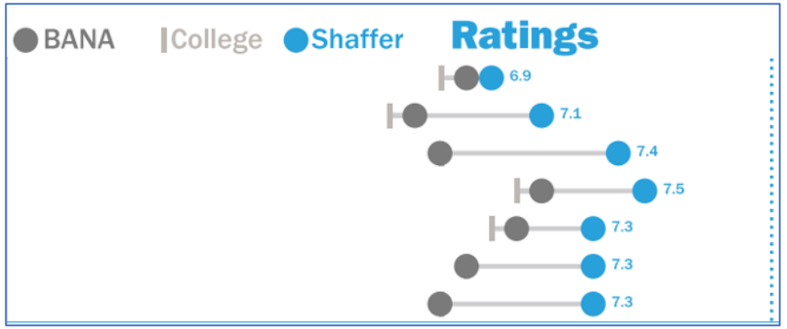
Fonte da Imagem: WEXLER, S., **SHAFFER**, J., & COTGREAVE, A. (2017)

Decisões de Design

- Barras de mesma largura



- Rotular somente uma das séries

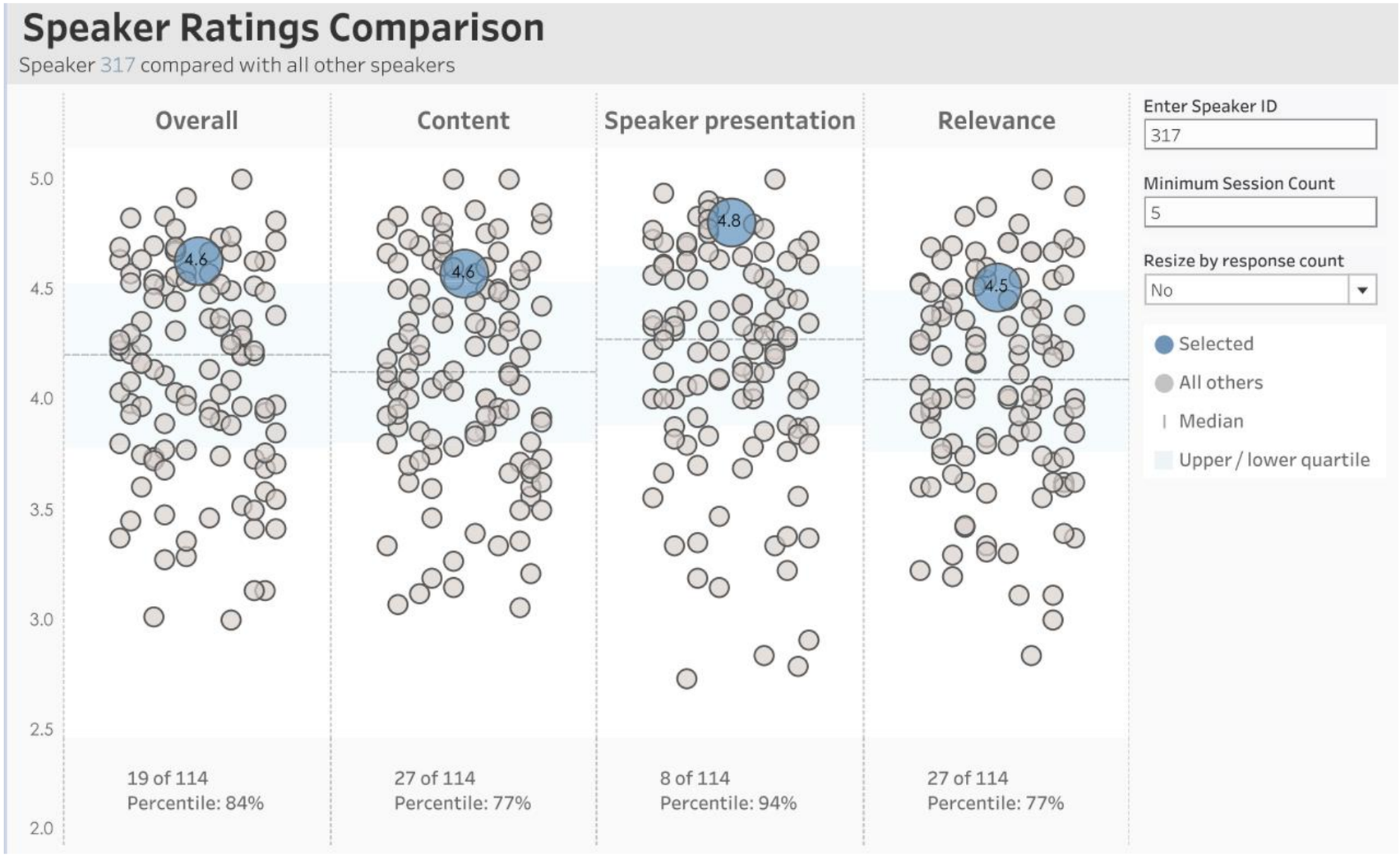


STEVE: Jeff's dot plot has become my go-to approach for comparing aggregated results from multiple sources (in this case an individual compared to a peer group compared to the college as a whole).

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

Dashboard de Oradores (jitterplot)

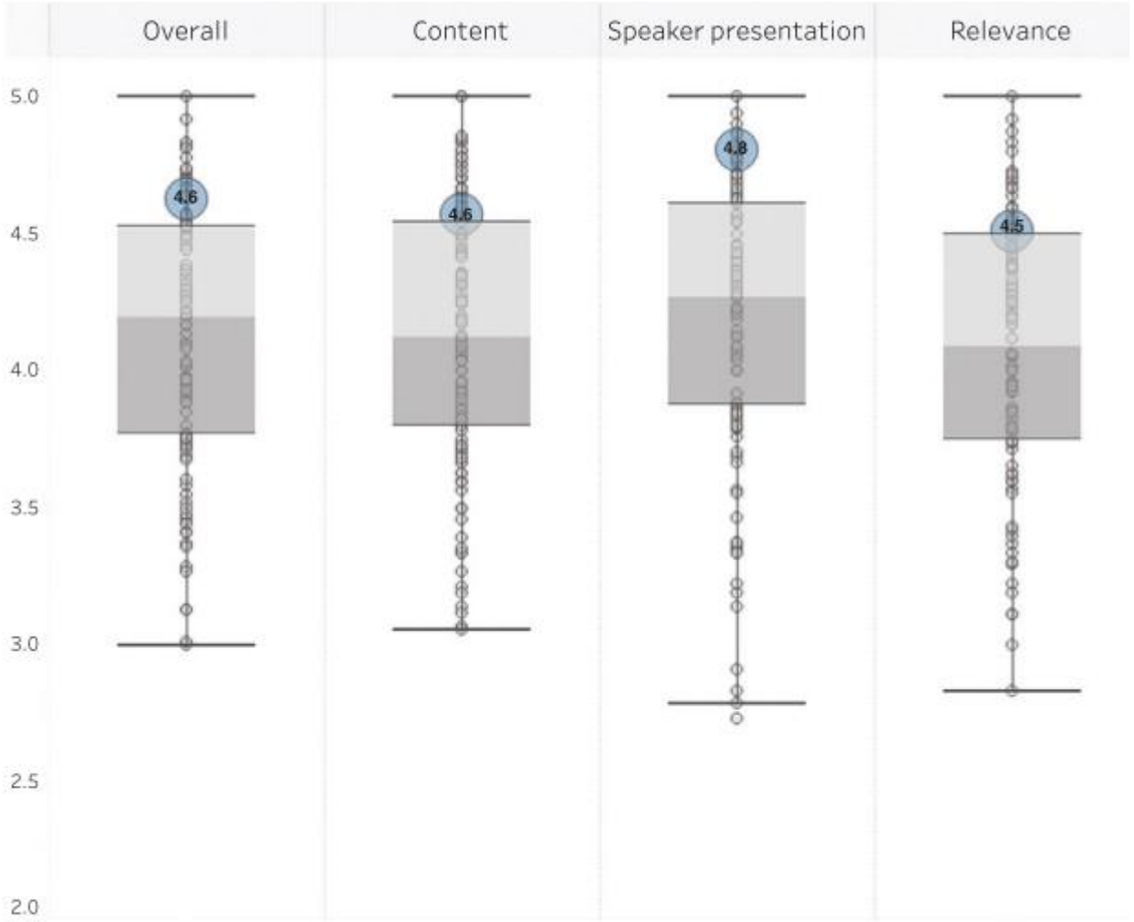
- A quantidade de pontos ajuda a contar uma história



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

Dashboard de Oradores

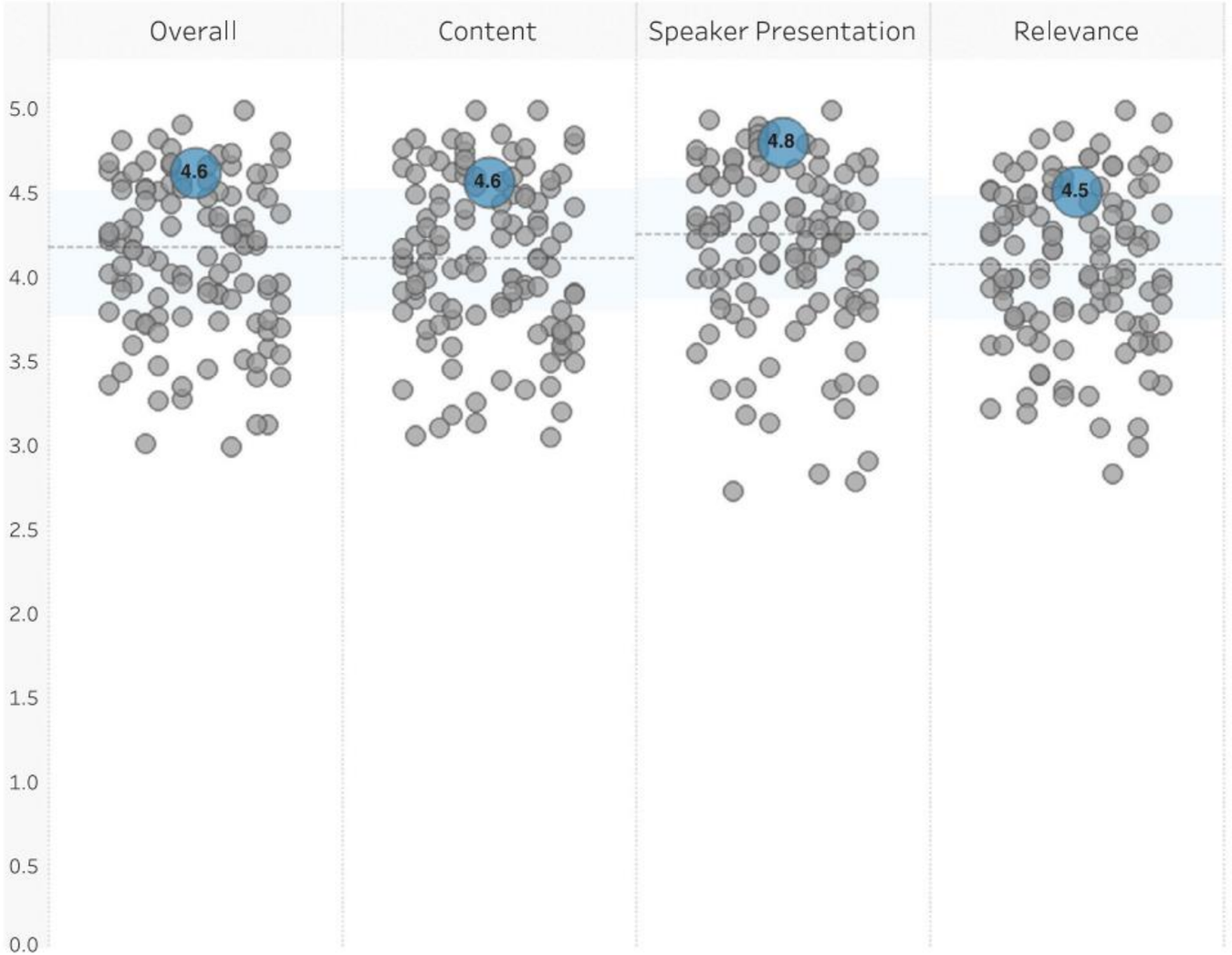
- Qual a desvantagem desta representação ?



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

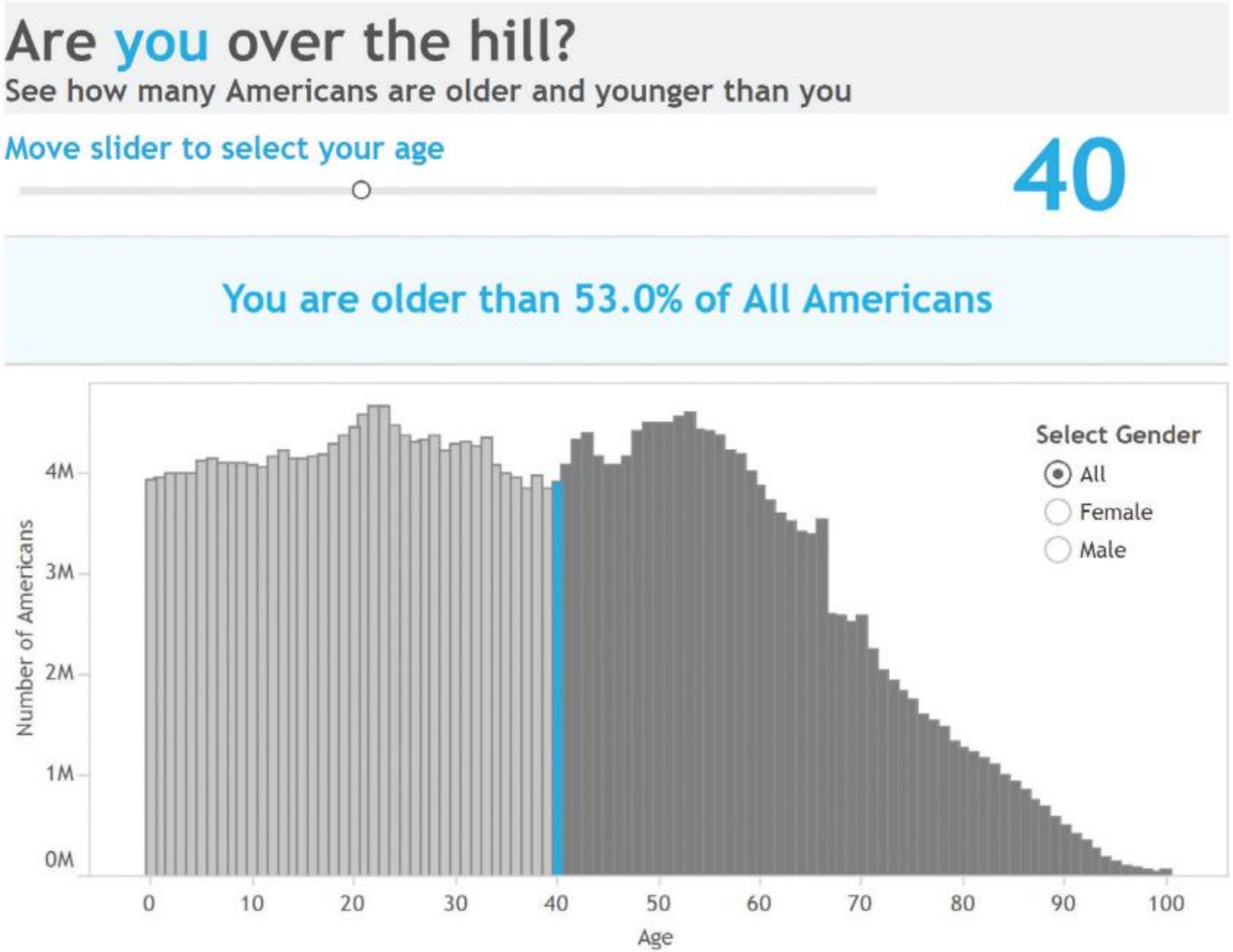
Dashboard de Oradores

- 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

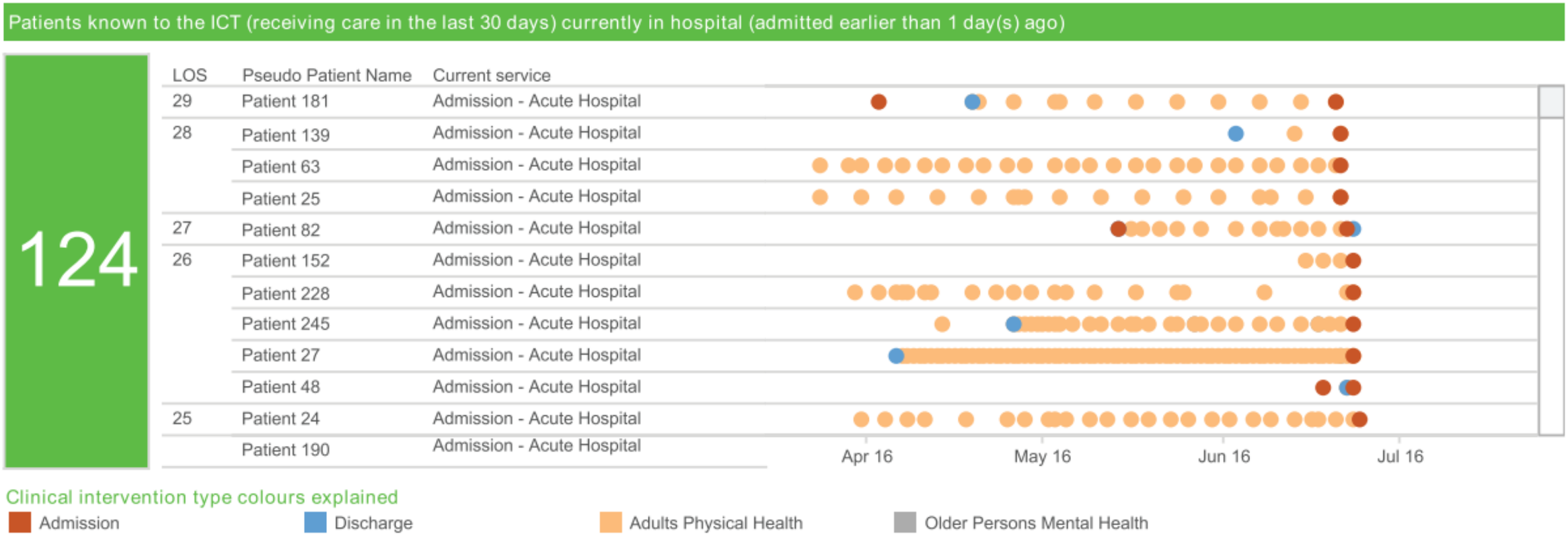
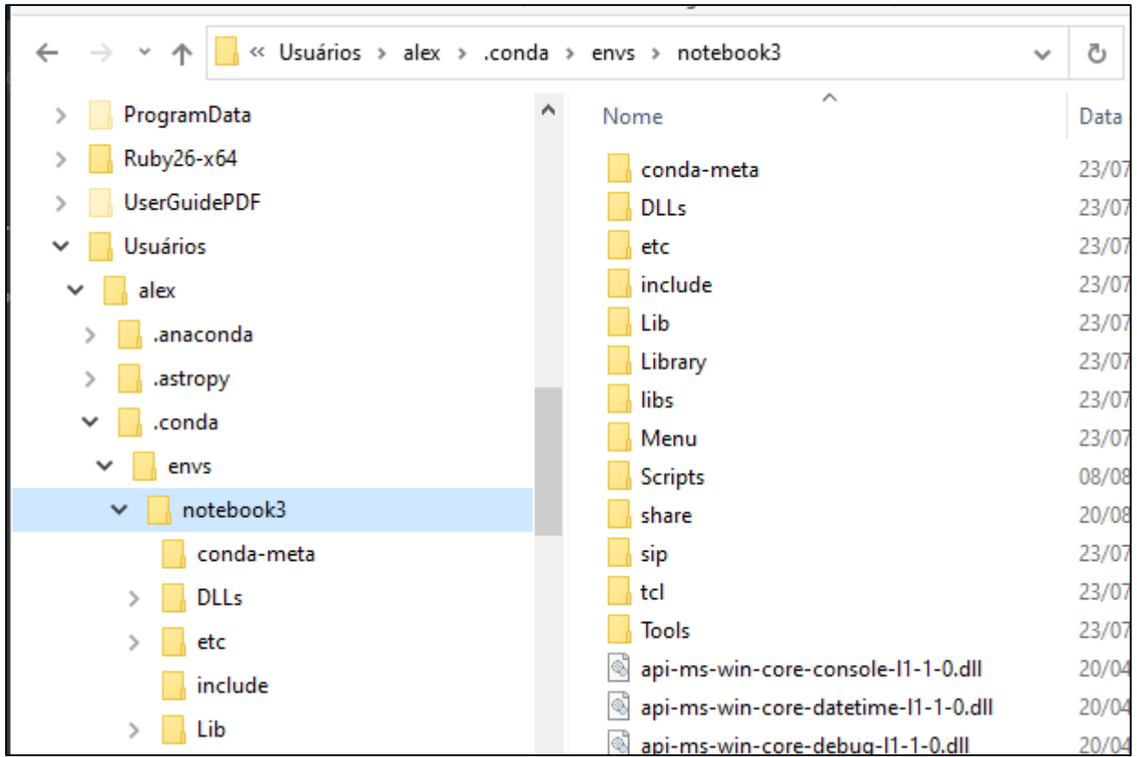


FIGURE 14.2 Patients admitted more than one day ago.

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

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

Criar projeto novo (vazio) no Pycharm

- Criar um novo projeto e um ambiente virtual pra ele
  - Demonstração
- Crie um arquivo teste.py, adicione uma função soma e teste-a
  - Demonstração
- Depure o seu código
  - Crie um breakpoint e execute o arquivo em modo debug
    - ✓ Inspecione as variáveis
    - ✓ adicione uma expressão aos Watches
    - ✓ Avance uma linha com o botão step over
    - ✓ Aprenda a usar os outros botões da barra de debug
  - Demonstração
- Instalação de pacotes pela interface do Pycharm
  - Demonstração

Criando um projeto no pycharm a partir de um repositório Git – Pré-requisito para a próxima aula

- VCS -> Get from Version Control (ou na janela de boas vindas)
  - No campo URL preencha a URL do repositório git deste curso
    - ✓ https://github.com/alexlopespereira/enapespcd2021.git
      - Isso vai fazer o clone do seu repositório
  - Demonstração
- Crie um ambiente virtual (se não lhe for oferecido um automaticamente)
  - Demonstração
- Instale pacotes descritos no arquivo requirements.txt
  - `pip install -r requirements.txt`
  - Demonstração

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

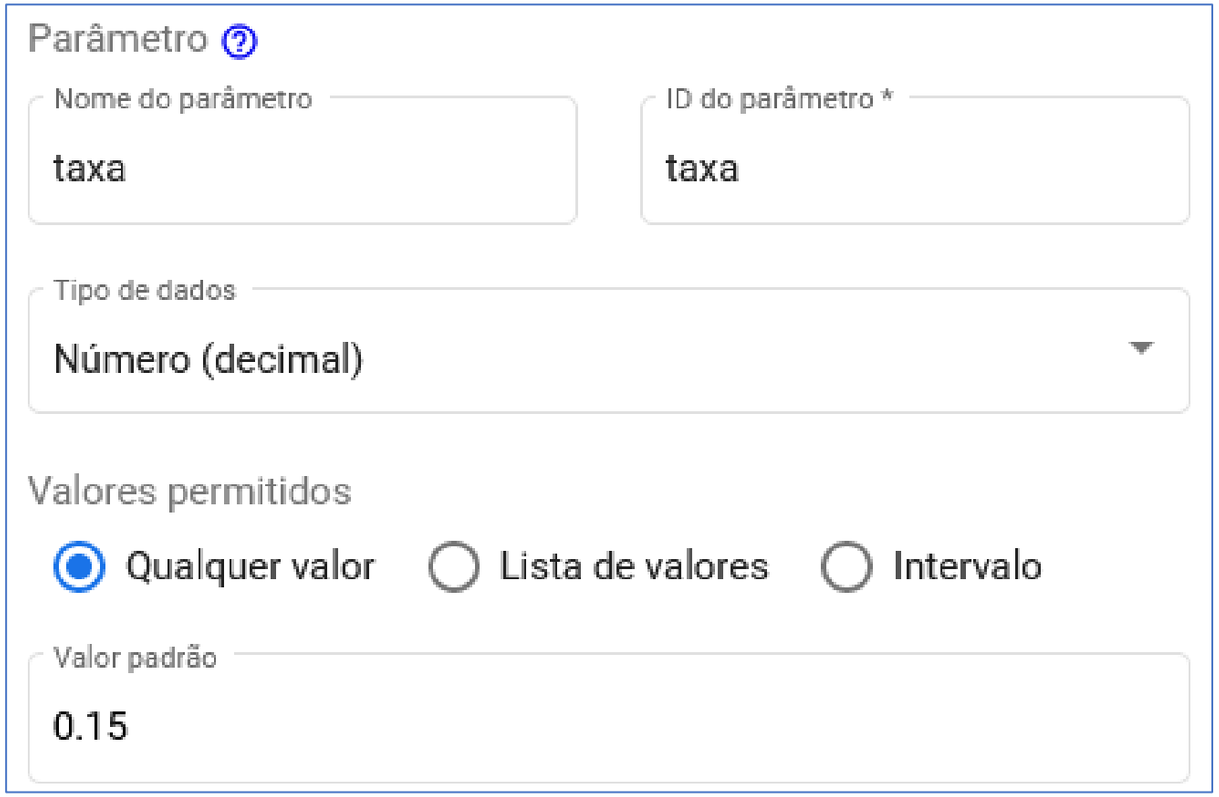
- Selenium é uma biblioteca de web scraping
  - Se você instalou os pacotes usando o comando pip (slide anterior)
    - ✓ O Selenium já está disponível para uso no seu projeto
- Ateste o correto funcionamento do Selenium com Debug (breakpoint)
  - Demonstração
    - ✓ Adicione um break point no script scrapy.py na linha que contem o código:
      - `driver.get(url)`
    - ✓ Execute o script scrapy.py em modo debug
      - Se o Selenium abrir uma janela do navegador,
        - Você conseguiu configurá-lo corretamente

Google Data Studio

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

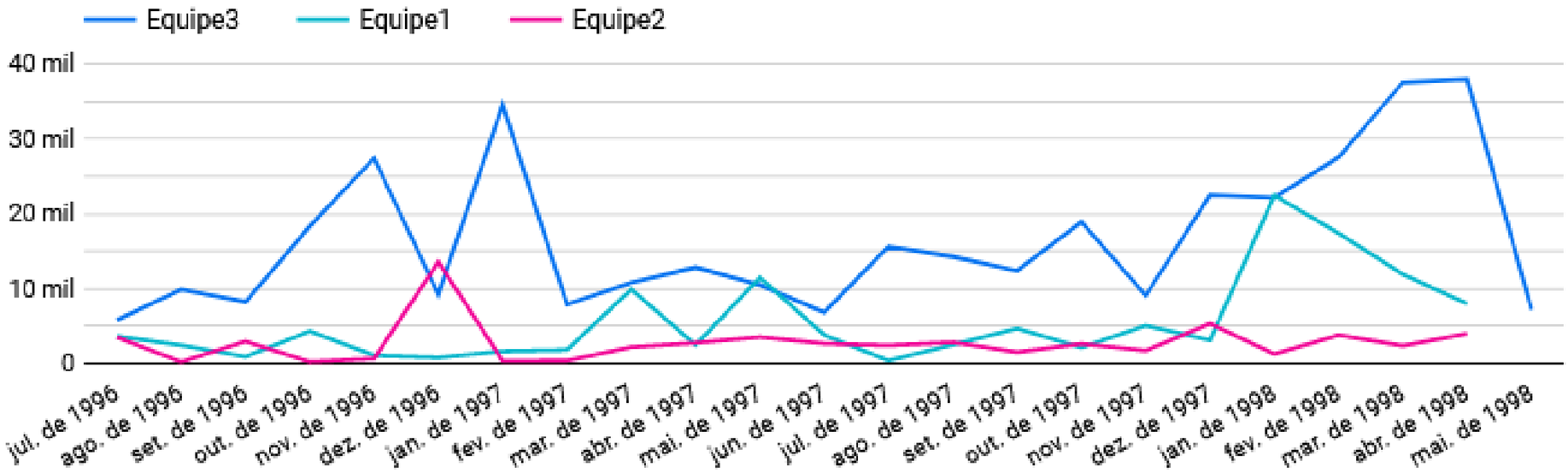
Parâmetro

- 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)

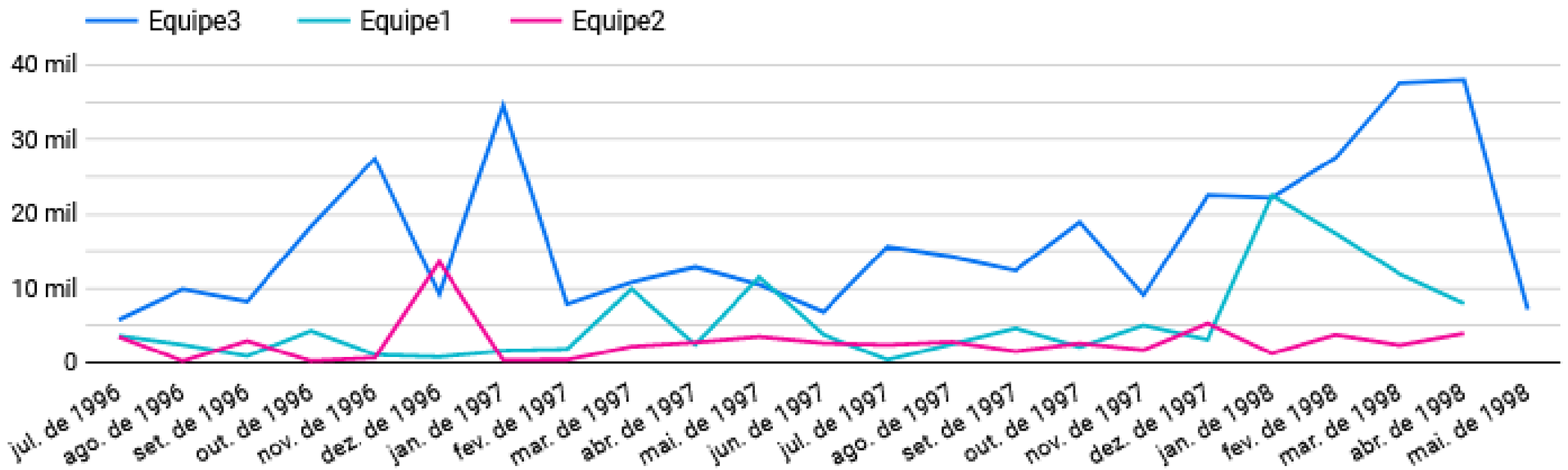
3.1) Criar um gráfico de linha do total de venda ($) por semana de cada equipe



- 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

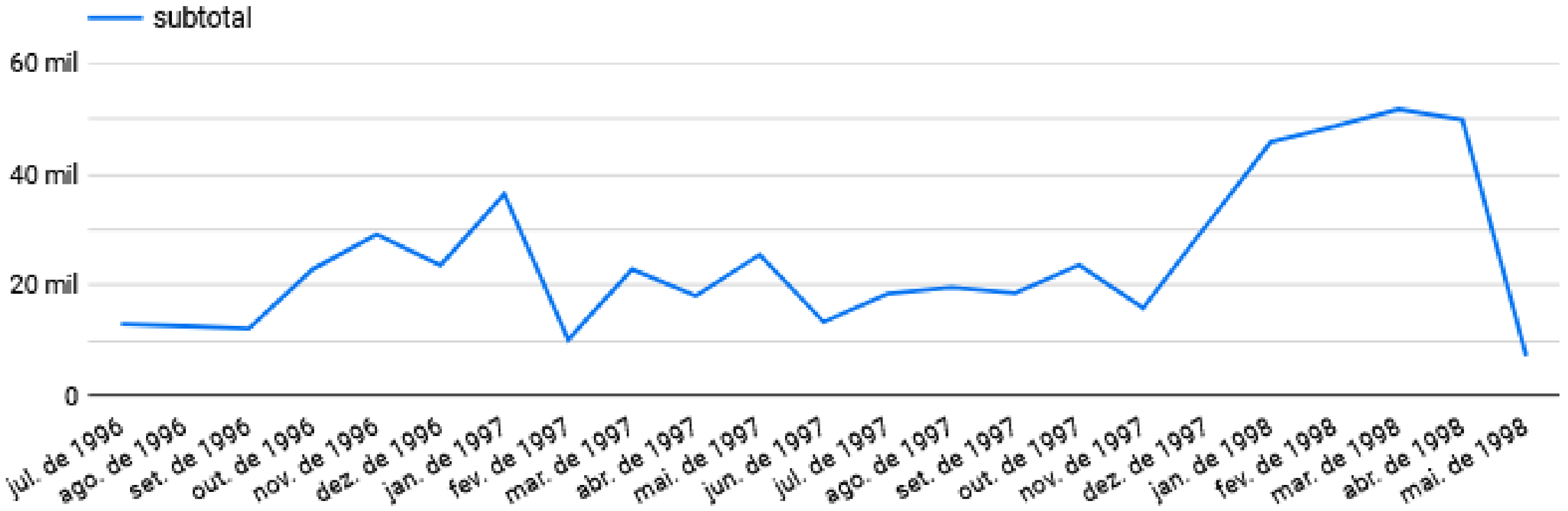
Removendo Clutter (Desordem)

	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 Data Studio,
  - Contendo informações sobre a população e o PIB dos municípios
    - ✓ Uma **tabela** com informações básicas
      - UF, Nome do Município, População, PIB e PIB Percapita
    - ✓ Um **gráfico** de linha com o PIB percapita dos estados de SP, RS, CE e AM
    - ✓ Uma **tabela** de heatmap com o PIB percapita dos estados ao longo dos 10 últimos anos
    - ✓ Um **dropdown** (lista suspensa) de filtro com a lista de UFs
  - **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

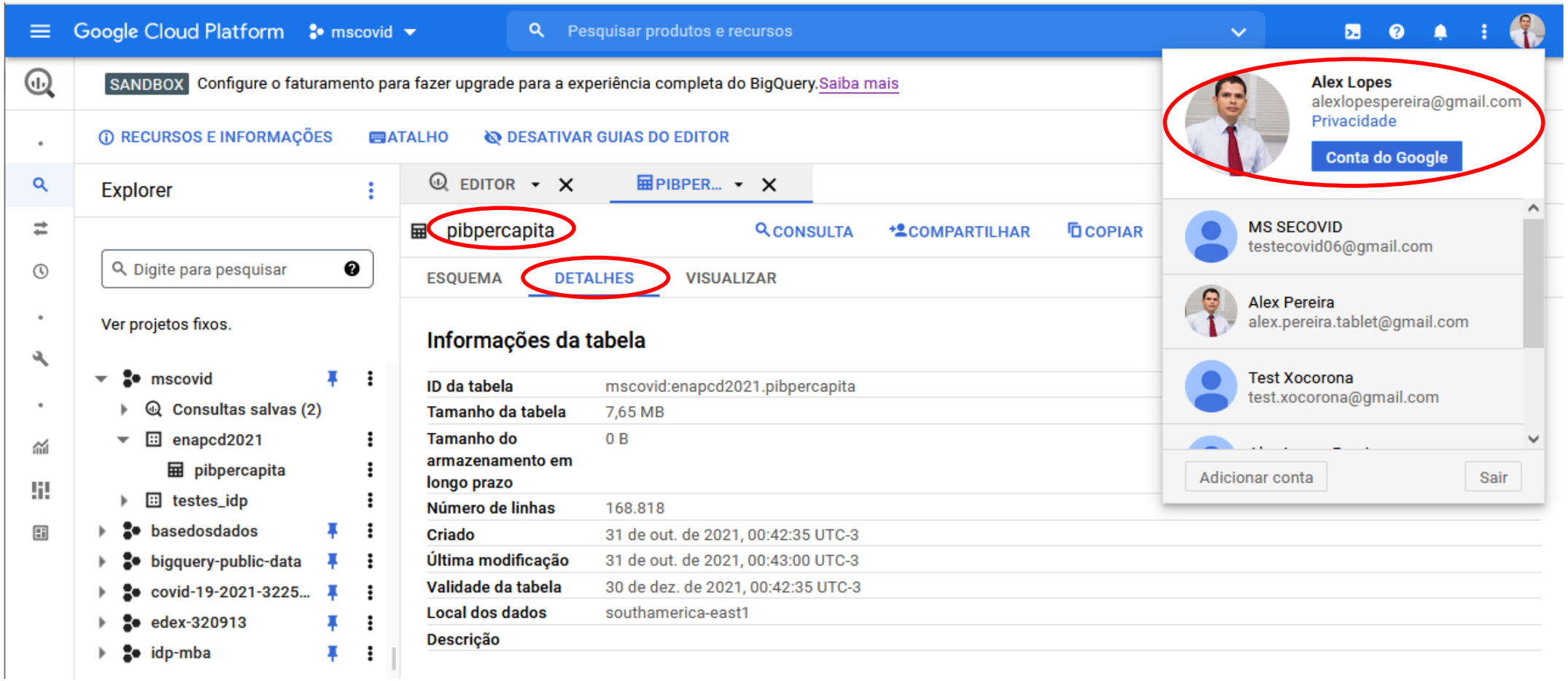
Exercício 3.1

- 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, e o link **público** para o seu caderno Colab.
    - ✓ **Não** compartilhe com meu email
    - ✓ Teste numa aba anônima
  - Demonstração

Link privado = não entregue / atrasado

Exercício 3.1

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



Exercício 3.2

- Crie uma proposta de visualização para o desafio de novembro
  - Makeover Challenge
    - ✓ da Cole Nussbaumer
- Submeta um print e uma contextualização do seu makeover
  - no mesmo link do Exercício 1.
    - ✓ E opcionalmente no site do desafio.