

Introdução a Ciência de Dados



Professor: Alex Pereira

O que estas séries de número te dizem sobre o respectivo fenómeno estudado ?

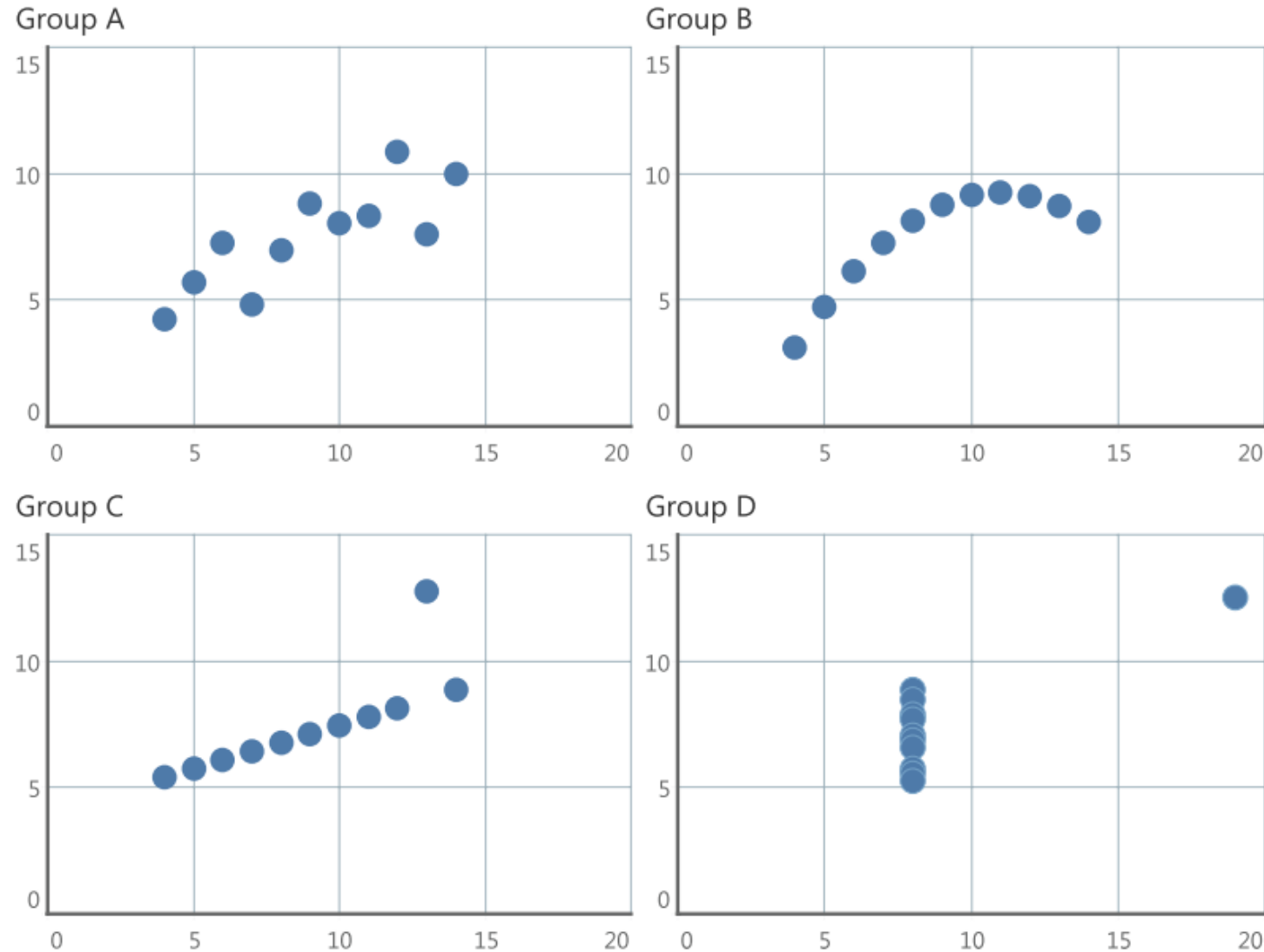
- Quarteto de Ascombe

- As 4 séries têm média e variância quase idênticas.

TABLE 1.1 Table with four groups of numbers: What do they tell you?

Group A		Group B		Group C		Group D	
x	y	x	y	x	y	x	y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Visualize a forma da sua série de dados antes de aferir suas conclusões



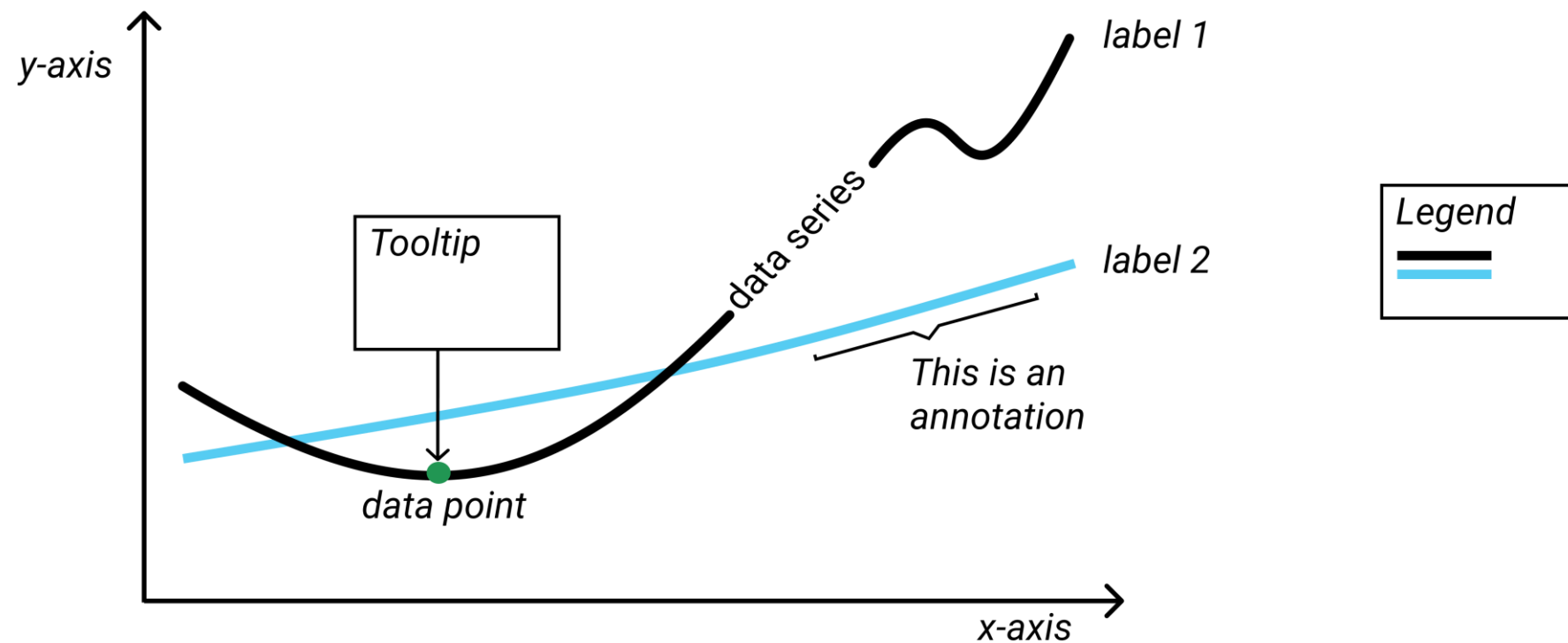
Definindo alguns conceitos importantes

- Título do Gráfico
 - Texto que resume o gráfico
 - ✓ Boa prática: frase de conclusão/ação em vez de frase descritiva
- Legenda
 - Referencia a aparência de uma medição do gráfico
 - ✓ Associando a um texto (label/rótulo)
- Anotações
 - Observações textuais dispostas em pontos específicos do gráfico
- Labels (rótulos)
 - Texto descritivo em várias partes dos gráficos
 - ✓ Eixos, legendas, anotações

Definindo alguns conceitos importantes

Title

Subtitle

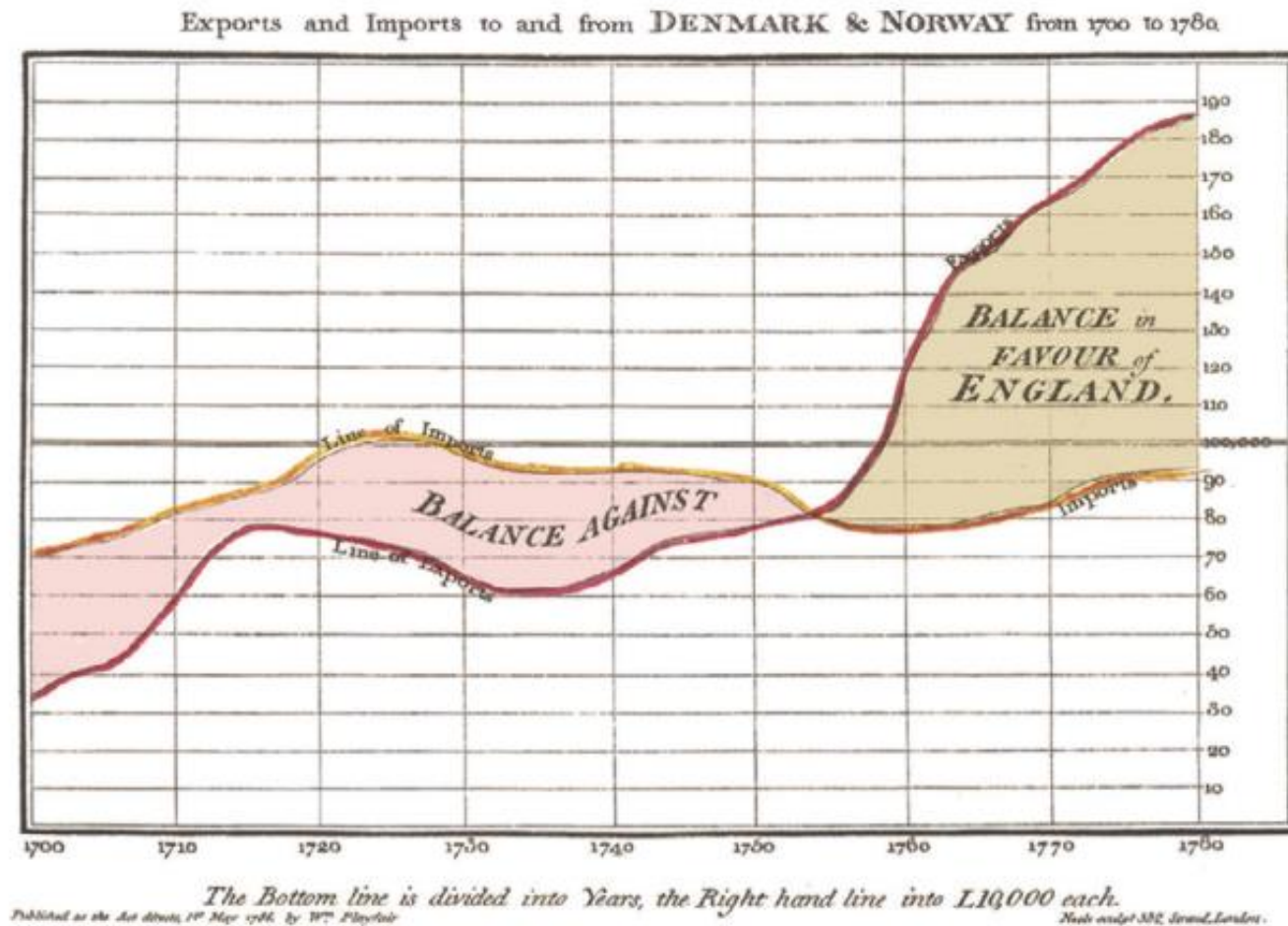


Notes/Source/Credits: _____

Análise temporal: além da linha do tempo

- Como complementar a análise de linha do tempo ?

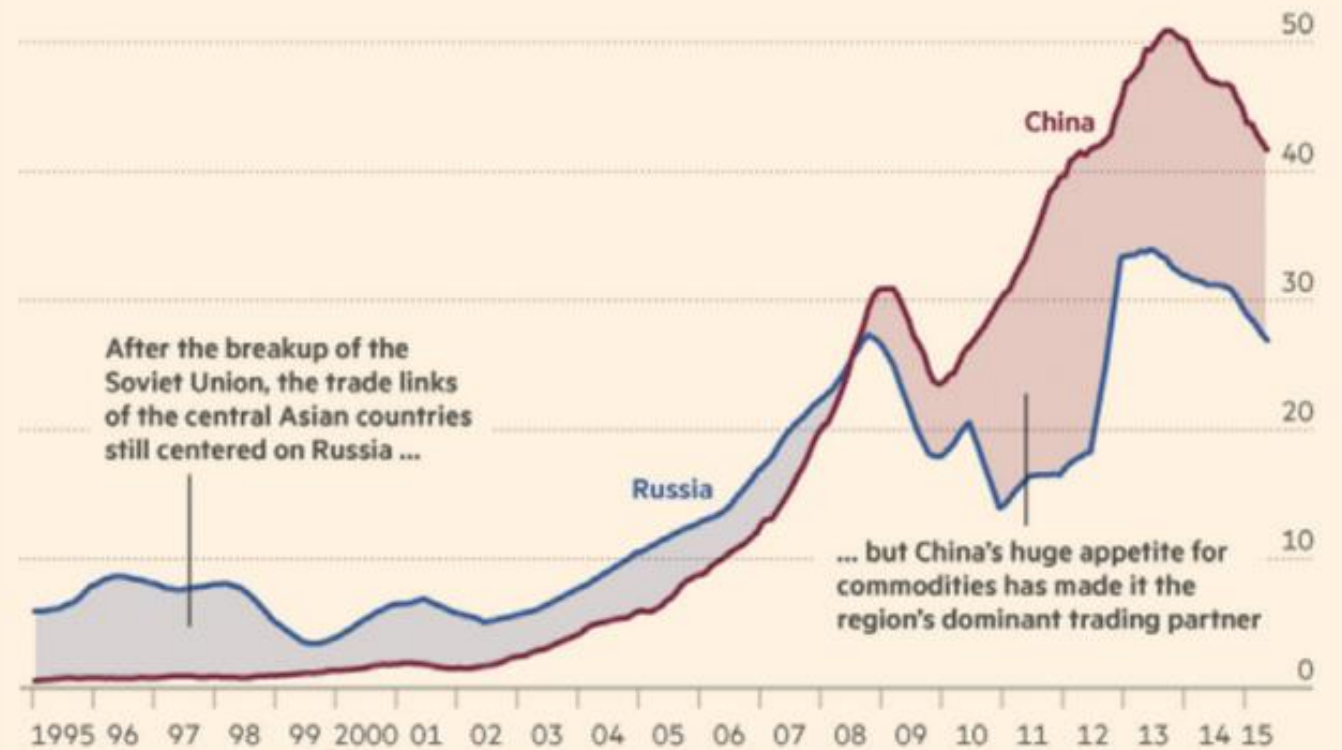
1786: William Playfair



2015: Financial Times

Trade with central Asian countries

Merchandise exports and imports, sum over previous 12 months (\$bn)



Source: IMF

FT

FIGURE 31.1 The first statistical timeline was drawn in 1786. Over 200 years later, we still use similar techniques.

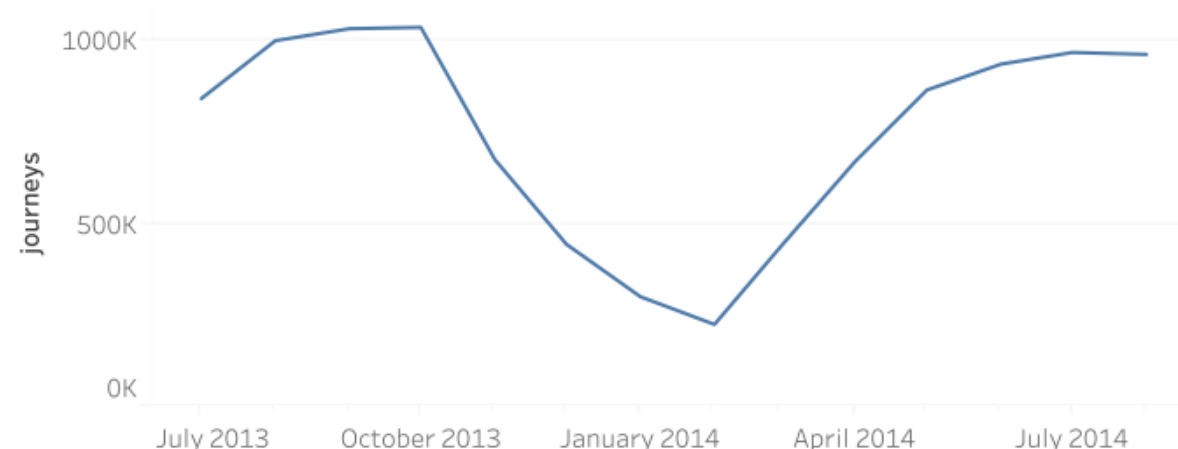
Perguntas que podemos fazer e analisar

1. Como o dia de hoje se compara ao início do período de análise?
2. Existem padrões cíclicos nos dados?
3. Como analisar tendências em duas dimensões de tempo?
4. Como ver o ranking, não o valor, ao longo do tempo?
5. Como comparar valores de coisas que não aconteceram ao mesmo tempo?
6. Como mostrar a duração de um evento?
7. Como me concentrar nos gargalos de um processo?

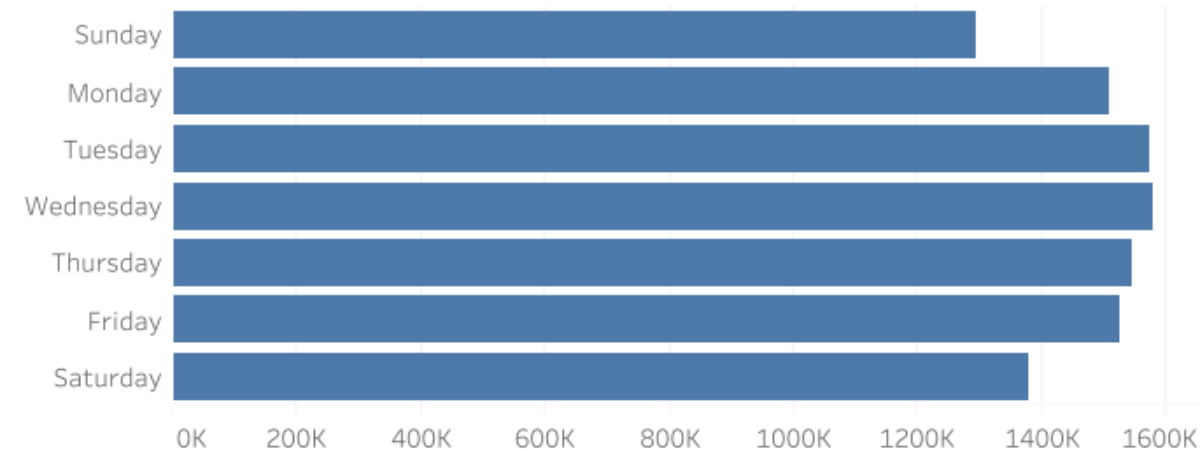
Algumas visualizações das jornadas de bicicleta em NY

Journeys on Citi Bike share program, New York

A: Monthly



C: By day of week



B: Daily



D: By hour of day

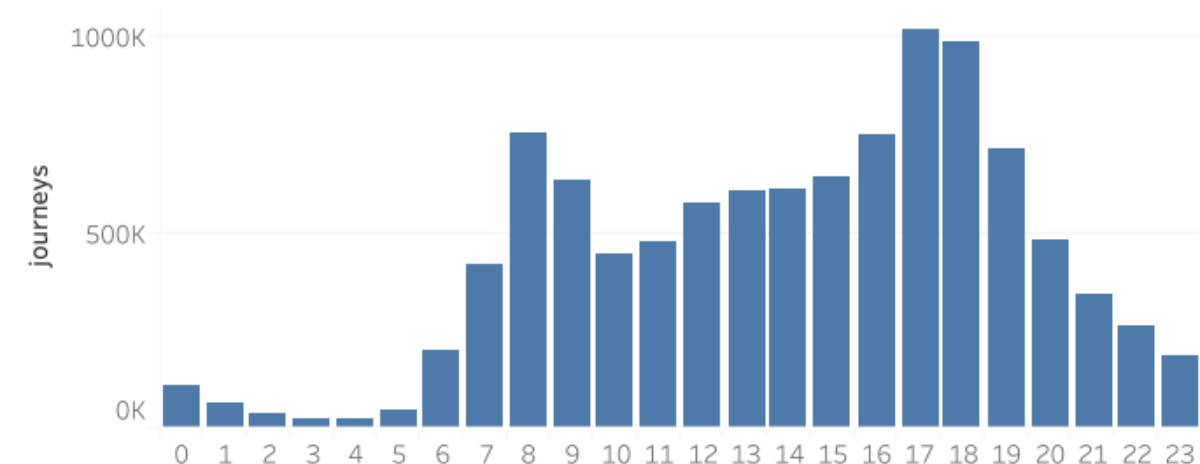
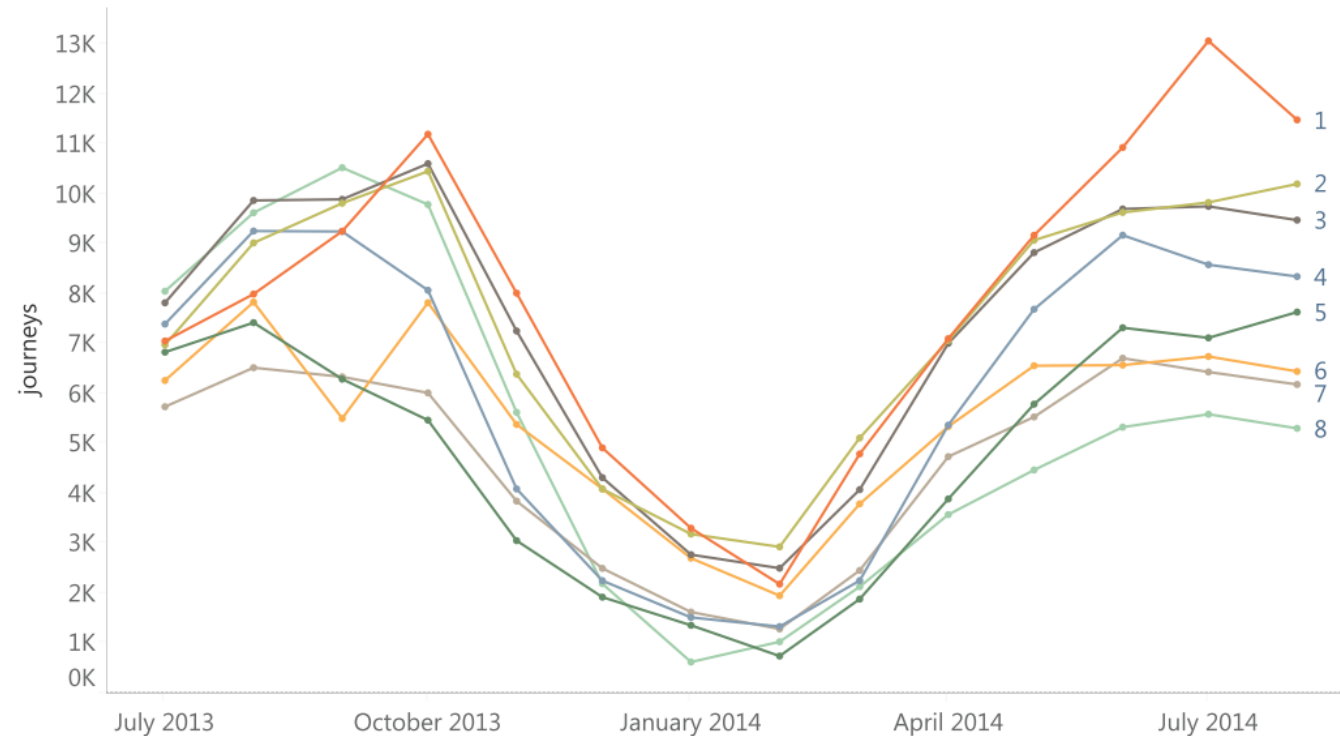


FIGURE 31.2 Four ways of showing Citi Bike journeys.

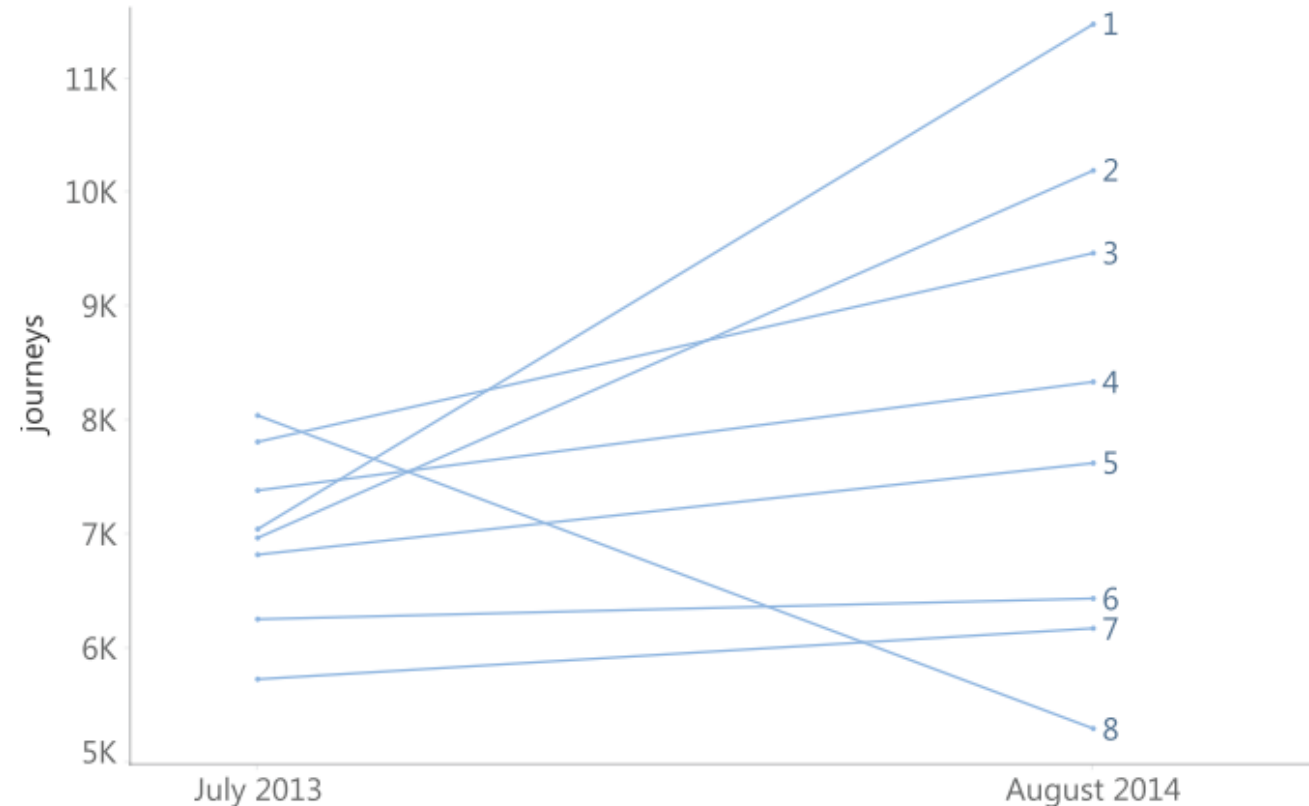
Como o dia de hoje se compara ao início do período de análise?

- Usar um slope chart (gráfico de inclinação)
 - Em vez de uma linha do tempo
 - ✓ Atenção para ruídos/oscilações momentâneas

The Top 8 Citi Bike Stations: Which stations have seen the biggest change in use since July 2013?



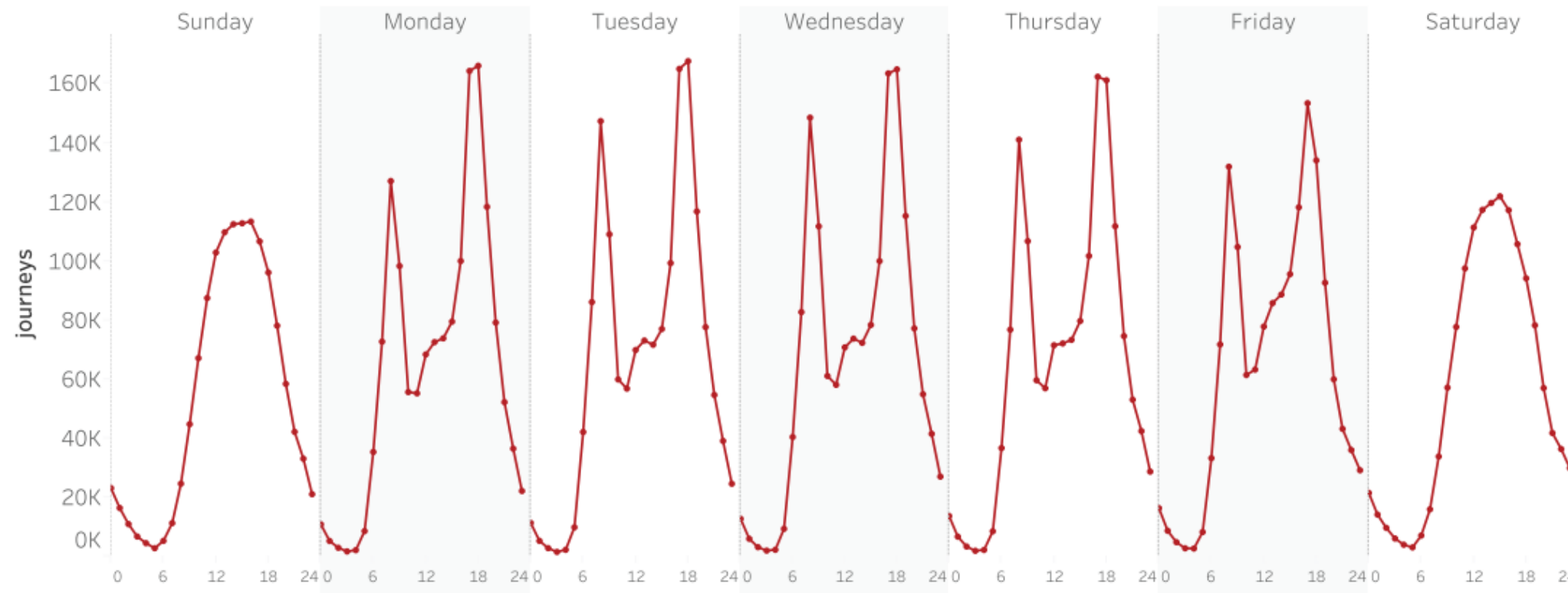
The Top 8 Citi Bike Stations: Which stations have seen the biggest change in use since July 2013?



Existem padrões cíclicos nos dados?

- Uma linha do tempo mostra que existem vales e picos
 - Mas pode não ser a melhor solução para compará-los
 - ✓ Qual dia tem maior demanda ao meio dia?

Citi Bike in New York: What do the hours of 8 to 9 a.m. and 12 to 1 p.m. look like?

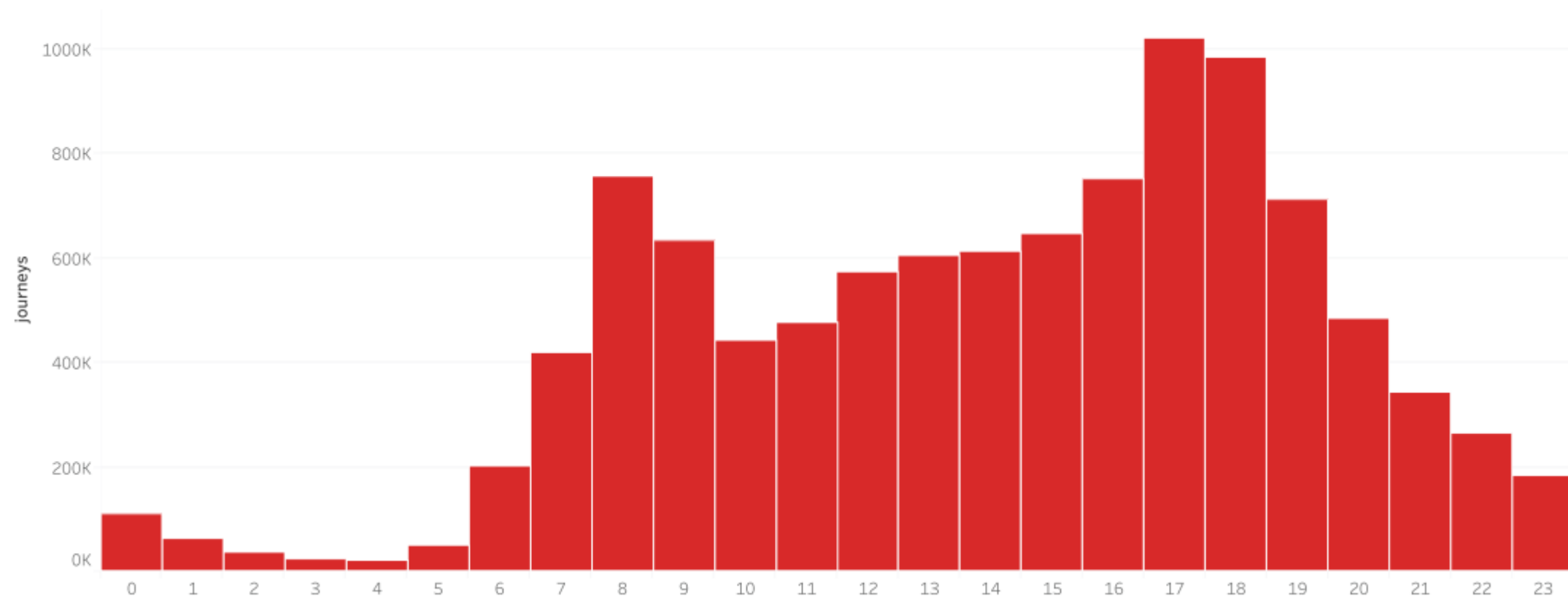


Fonte: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Existem padrões cíclicos nos dados?

- O gráfico de barras é bom para comparar a média
 - em cada hora (categoria)
 - ✓ Mas não é útil para comparar valores dentro de cada hora (categoria)

Citi Bike in New York: What do the hours of 8 to 9 a.m. and 12 to 1 p.m. look like?



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

Qual dia tem maior demanda ao meio dia?

- Num cycle plot você encontra a resposta rapidamente
 - A linha preta horizontal representa a média dos 7 dias

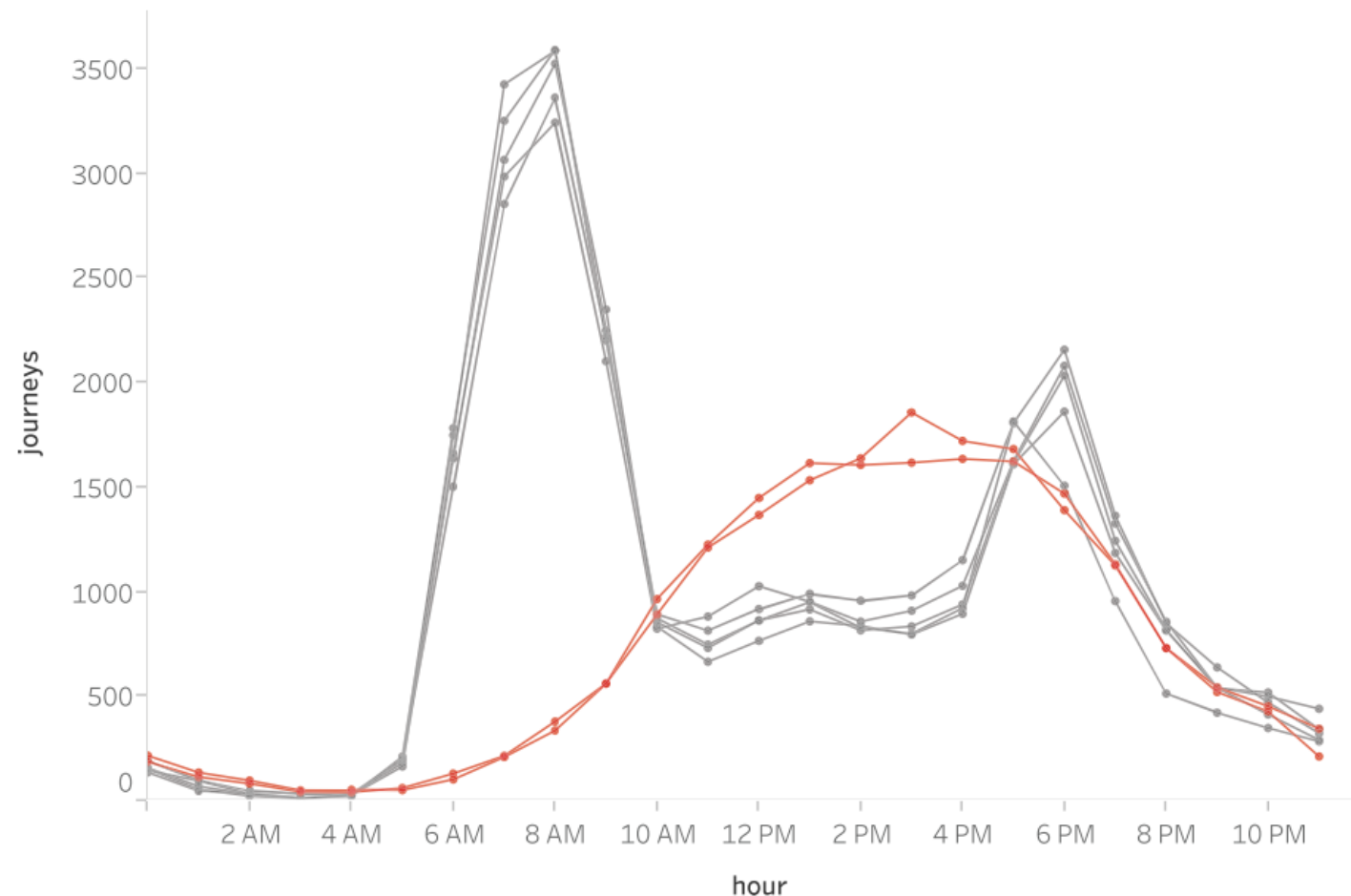
Citi Bike in New York: What do the hours of 8 to 9 a.m. and 12 to 1 p.m. look like?



Se o espaço horizontal for uma restrição

- Use um gráfico de linha temporal com várias séries

Citi Bike in New York: Journeys by hour of day and day of week
(weekend days are red, weekdays are grey)



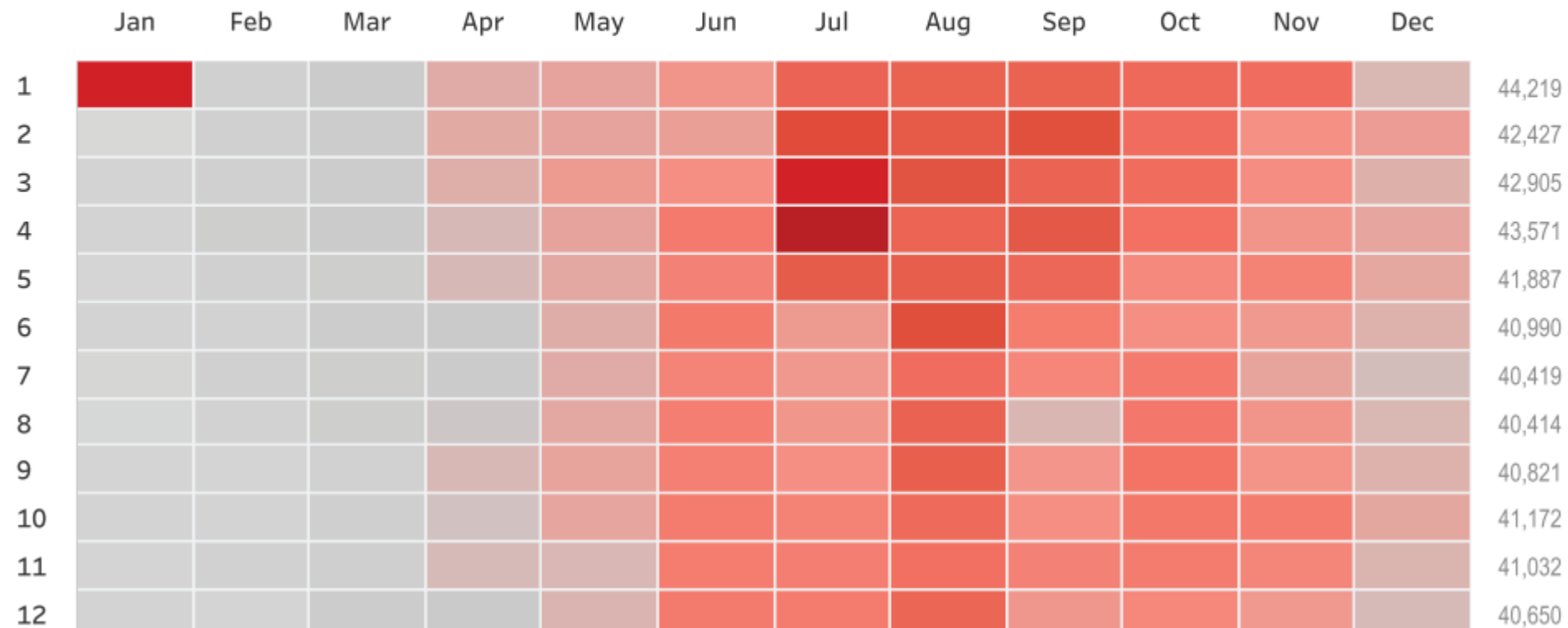
Fonte: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Como analisar tendências em duas dimensões de tempo?

- Com um heatmap (mapa de calor)

US Road Fatalities by month and day,
1975-2011

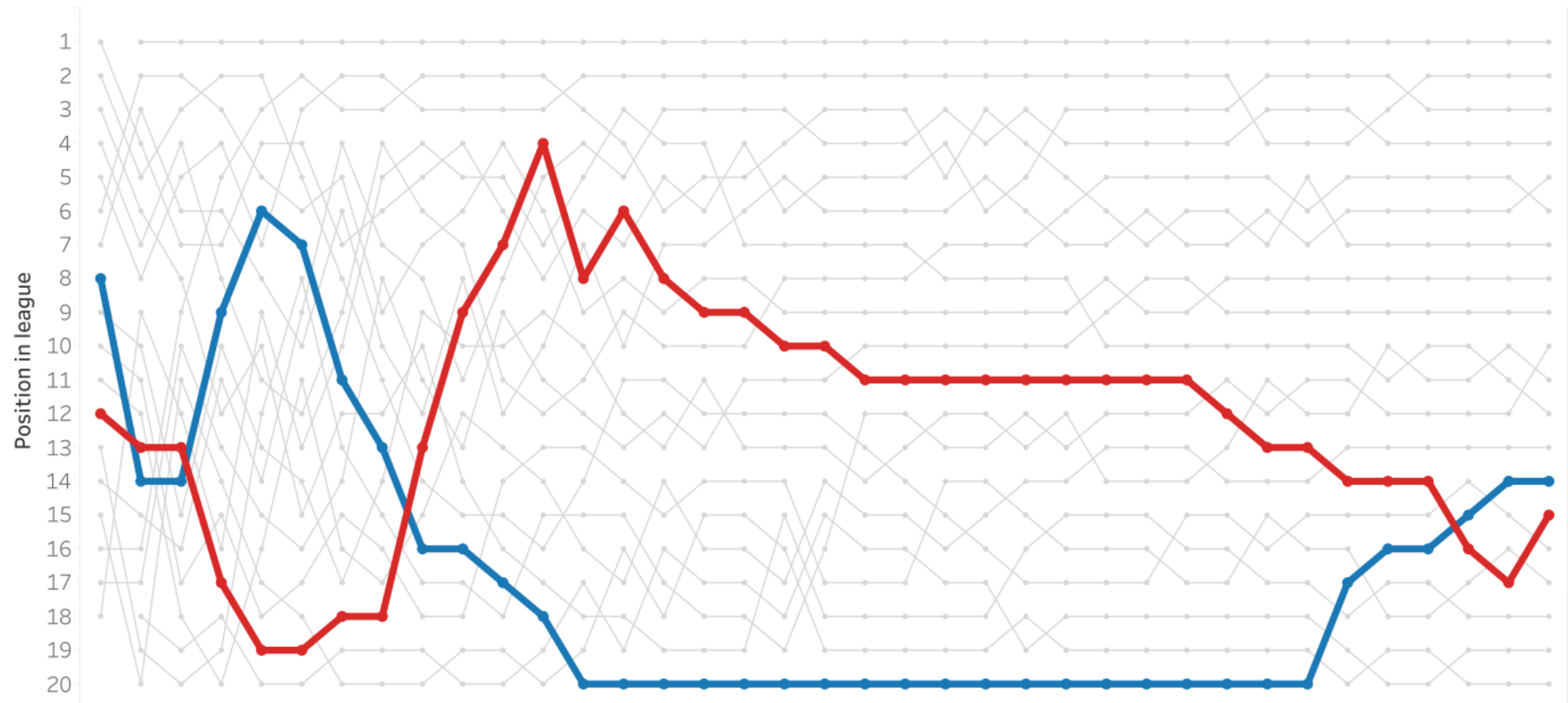
Fatalities
800 5,200



Como ver o ranking, não o valor, ao longo do tempo?

- Com um bump chart
 - Em alguns contextos o ranking importa mais do que o valor exato.

English Premier League 2014/2015: How did **Leicester** and **Newcastle** perform?

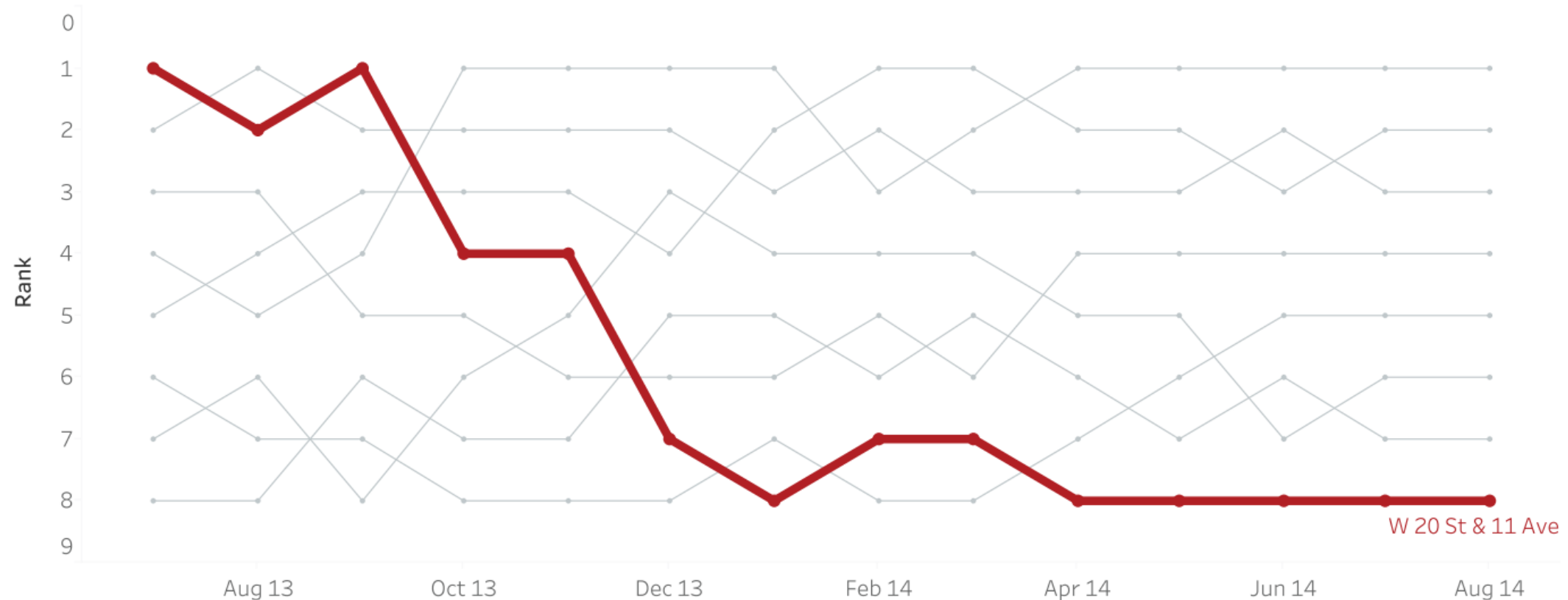


Fonte: WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017)

Como ver o ranking, não o valor, ao longo do tempo?

- Com um bump chart
 - Em alguns contextos o ranking importa mais do que o valor exato.

Citi Bike's Top 8 stations: How did W 20 St & 11 Ave change in rank over time?

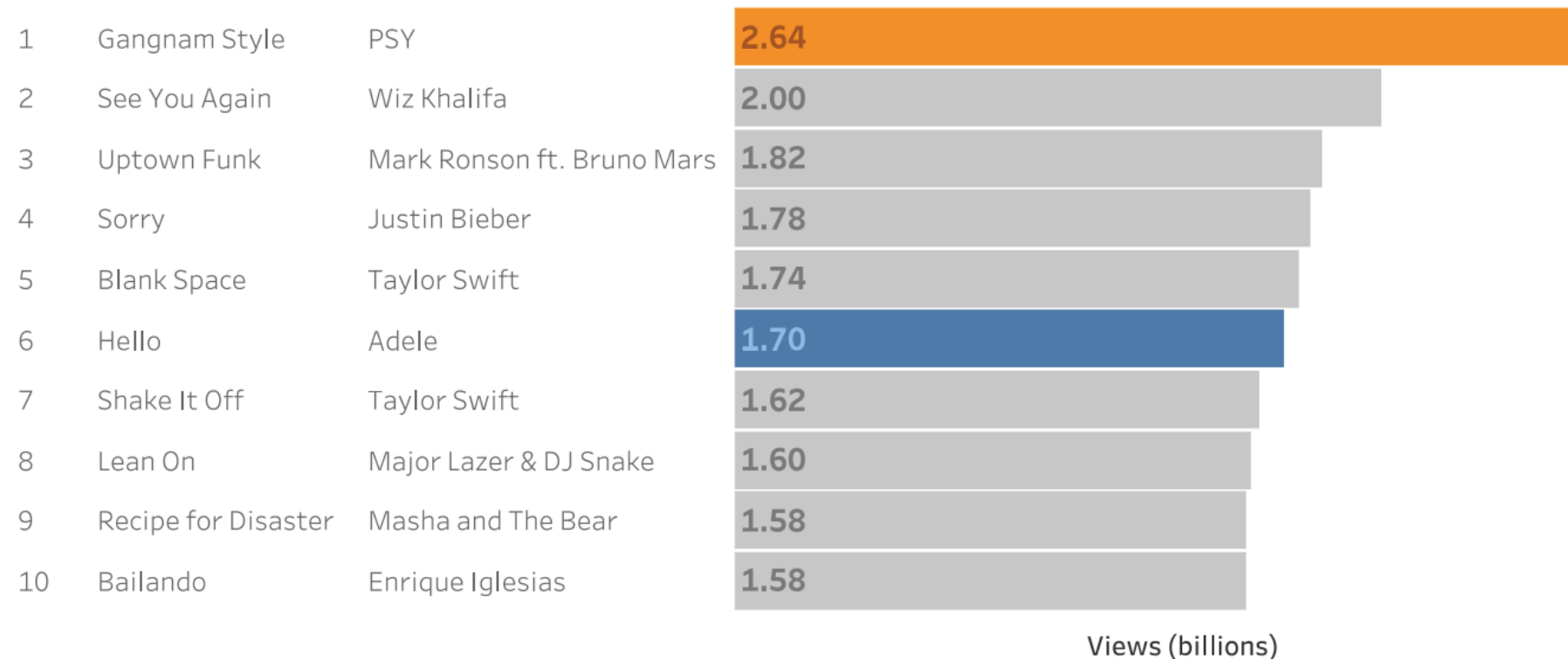


Como comparar valores de coisas que não aconteceram ao mesmo tempo?

- Qual vídeo viralizou (1 bilhão de visualizações) mais rápido ?
 - Um gráfico de barras pode ser o início da sua análise exploratória

The Top 10 most viewed videos on YouTube

Highlighted: **Hello by Adele** and **Gangnam Style by PSY**

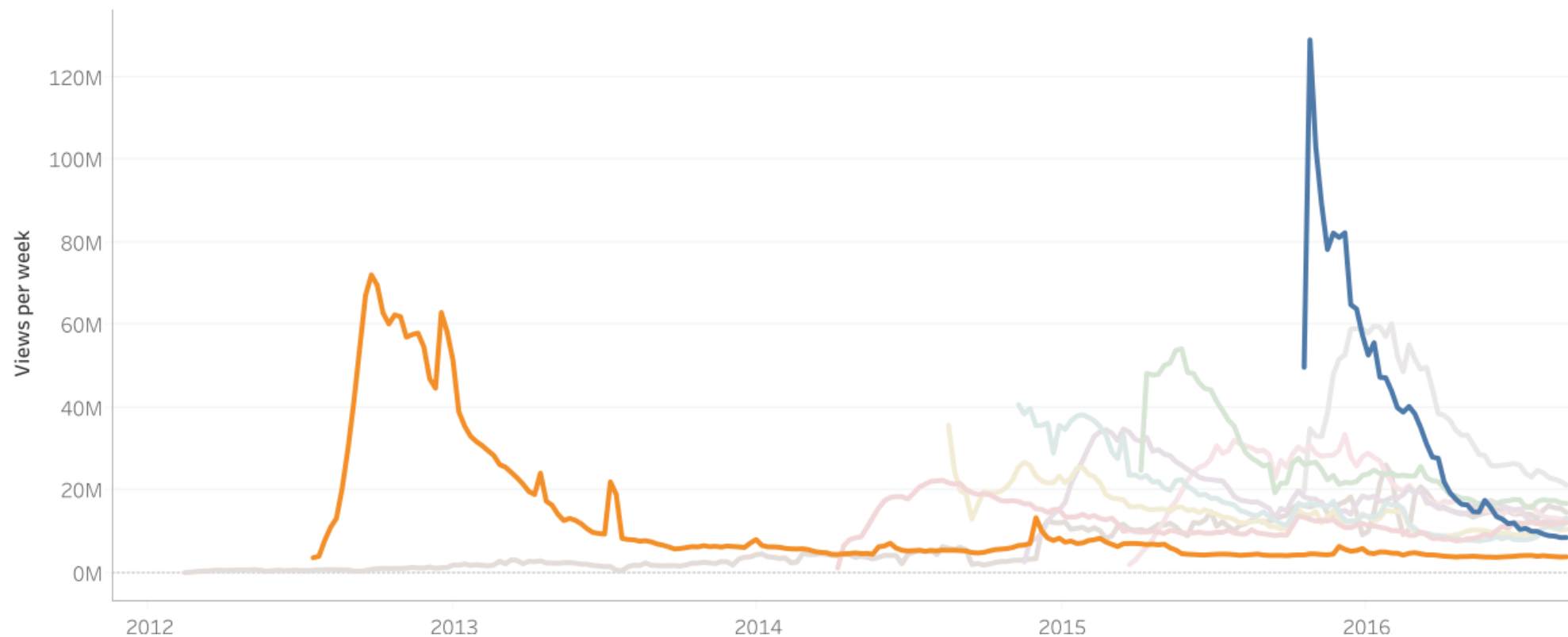


Como comparar valores de coisas que não aconteceram ao mesmo tempo?

- Qual vídeo viralizou (1 bilhão de visualizações) mais rápido ?
 - Um gráfico de linha dá um indício, mas não é contundente.

Views per week of the 10 most popular YouTube videos.

Highlighted: **Hello by Adele** and **Gangnam Style by PSY**

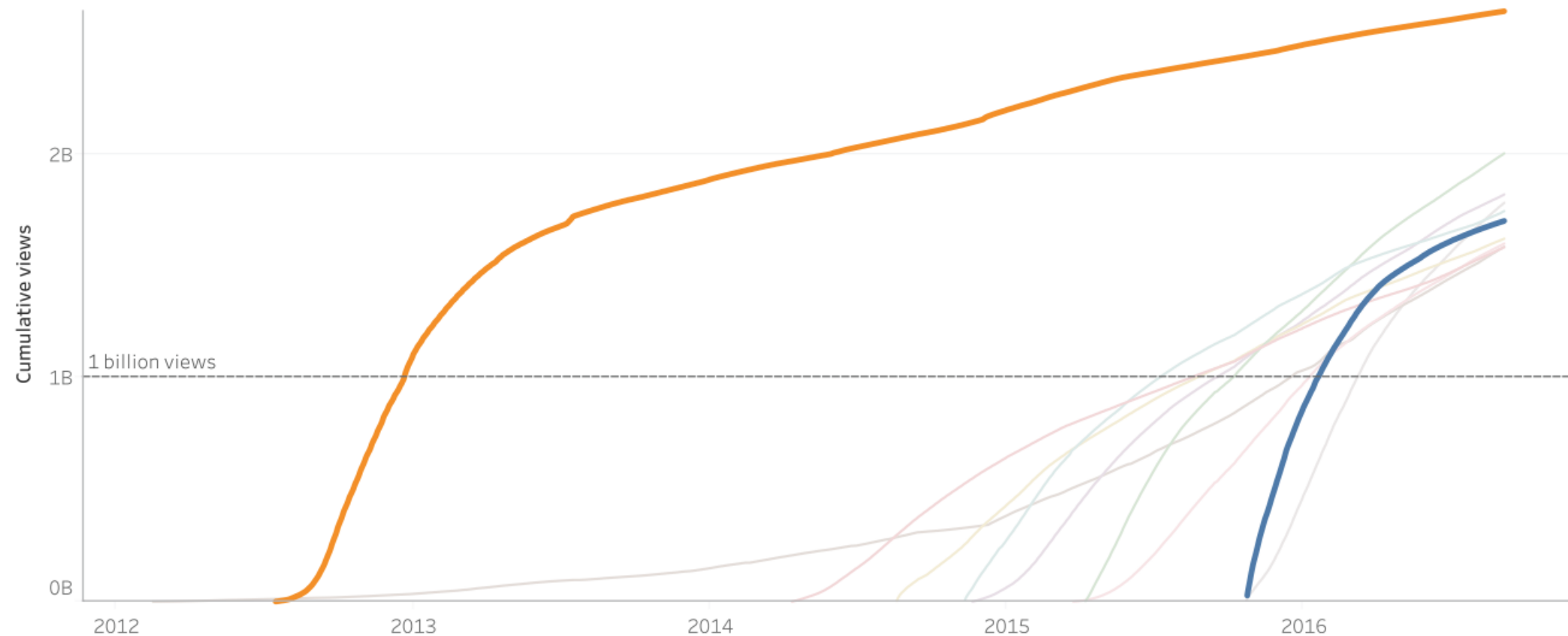


Como comparar valores de coisas que não aconteceram ao mesmo tempo?

- Qual vídeo viralizou (1 bilhão de visualizações) mais rápido ?
 - Um gráfico de linha acumulado é mais preciso

Cumulative daily views of the 10 most popular YouTube videos.

Highlighted: **Hello by Adele** and **Gangnam Style by PSY**

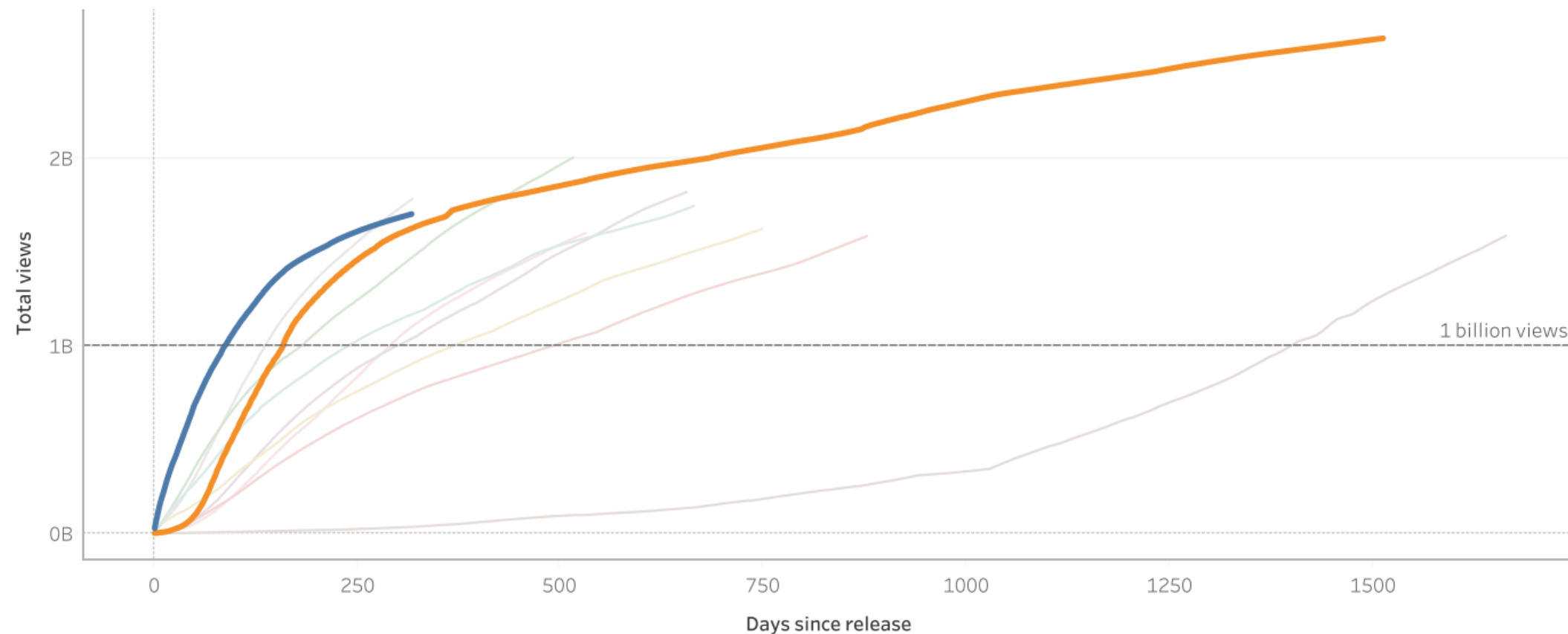


Como comparar valores de coisas que não aconteceram ao mesmo tempo?

- Qual vídeo viralizou (1 bilhão de visualizações) mais rápido ?
 - Um gráfico de linha acumulado com tempo relativo não deixa dúvidas

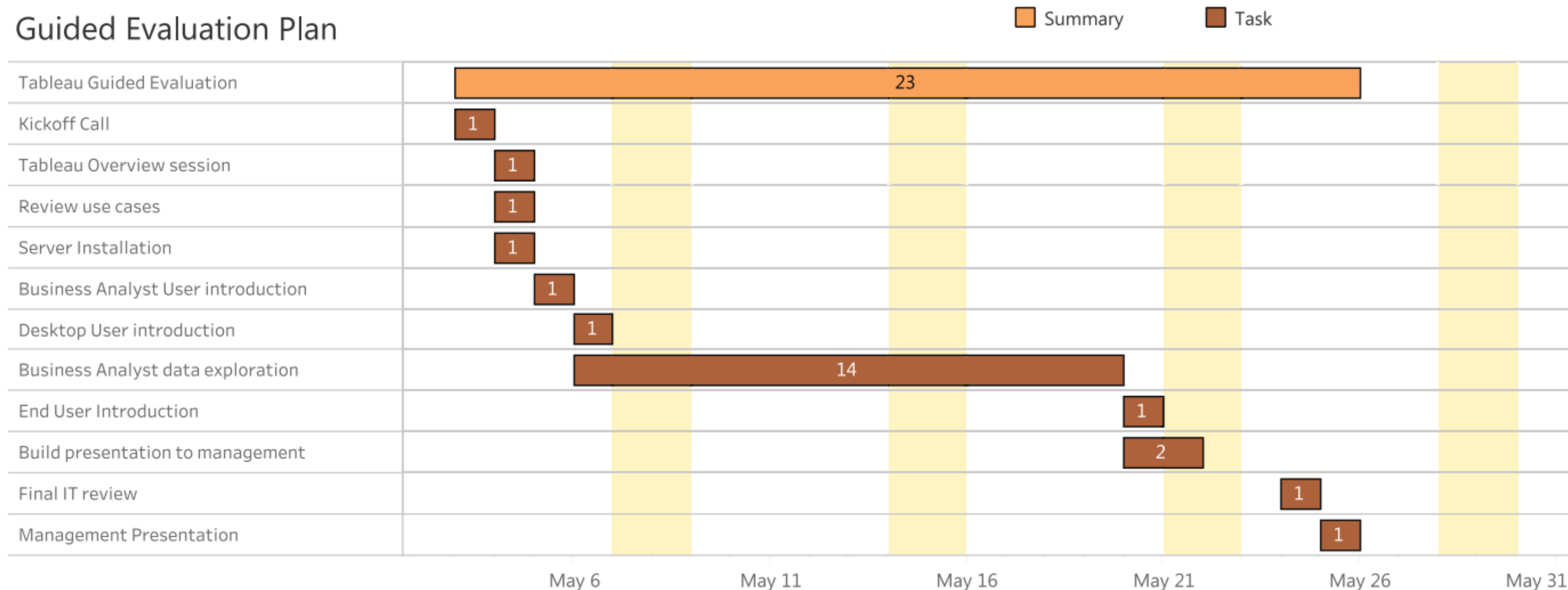
Cumulative daily views of the 10 most popular YouTube videos.

Highlighted: **Hello by Adele** and **Gangnam Style by PSY**



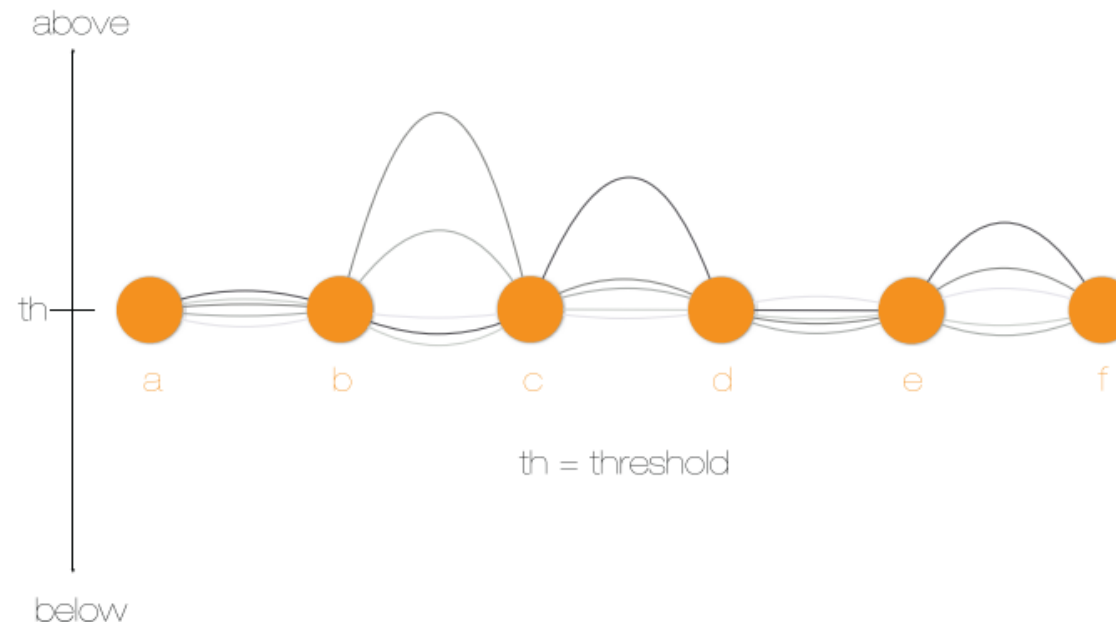
Como mostrar a duração de um evento?

- Um gráfico Gantt é uma ótima opção
 - inclusive para visualizar o caminho crítico de um projeto



Como me concentrar nos gargalos de um processo?

- Com um Jump plot
 - Os nós são eventos/etapas de um processo
 - ✓ dispostos sequencialmente da esquerda para a direita
 - A altura do arco é proporcional ao tempo transcorrido entre dois nós
 - ✓ Se houver um threshold diferente de zero, a altura será proporcional a diferença ao threshold



Projetos de Dashboard não têm fim

- Sua organização evolui com o tempo,
 - e seus painéis também deveriam acompanhar a evolução
 - ✓ do contrário, eles correm o risco de se tornarem painéis inertes.
- Fitbit do autor (Andy Gotgreave)
 - O aparelho e o dashboard mudaram seus hábitos
 - Com o passar do tempo as métricas estabilizaram
 - ✓ Os novos hábitos de atividade física se tornaram recorrentes
 - O dashboard permaneceu o mesmo (sem novidades)
 - ✓ O Andy não precisava mais do dashboard para saber as métricas
 - ✓ Mas queria saber outras informações
 - Recordes, variações com clima, humor e trabalho
 - Quando quebrou, não comprou outro Fitbit



Projetos de Dashboard não têm fim

- Desconhecido conhecido (Donald Rumsfeld)
 - Não sabemos a resposta, mas já nos interessamos e elaboramos as perguntas
 - ✓ Eventualmente respondidas num painel
- Desconhecido desconhecido
 - Nem se quer elaboramos perguntas ou nos chamou a atenção
- Análises quantitativas em geral, e dashboards
 - Fazem essa cadeia de ideias se movimentar
 - ✓ E ajudam a explorar e documentar os limites da atenção/consciência
 - Iterar é esperado e também uma medida de progresso
 - ✓ Minimize o tempo/custo de iterar!

Se entender, já está falando a língua dos nerds

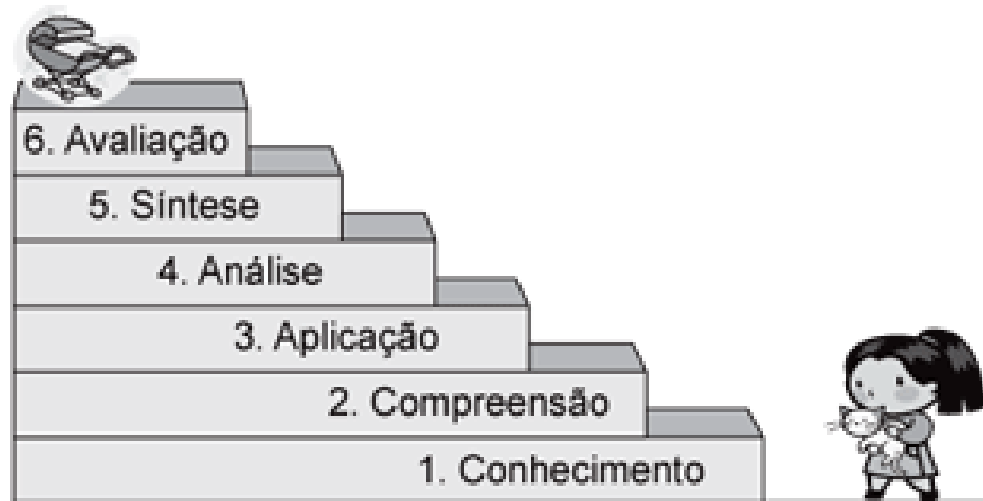


Figura 1. Categorias do domínio cognitivo proposto por Bloom, Englehart, Furst, Hill e Krathwohl, que ficou conhecido como Taxonomia de Bloom.



Se entender, já está falando a língua dos nerds



Avaliação Nacional da Alfabetização (ANA)

- Objetivo

- adicionar estas informações ao modelo de dados da semana anterior
 - ✓ Aquele com dados de população e PIB
 - Como fazer?

ano	sigla_uf	id_municipio	populacao	nome_municipio	pib
2002	RO	1100023	78039.0	Ariquemes	449592816.0
2003	RO	1100023	79680.0	Ariquemes	539636214.0
2004	RO	1100023	86901.0	Ariquemes	657193231.0
2005	RO	1100023	85031.0	Ariquemes	749021187.0
2006	RO	1100023	86924.0	Ariquemes	790696634.0

pibpercapita

Rede

- 1 = Federal
- 2 = Estadual
- 3 = Municipal
- 4 = Privada

ano	id_municipio	rede	id_escola	numero_validos_lp	numero_validos_mt	media_lp_escrita	media_lp_leitura	media_mt
2016	1100015	3	11024372	16	15	475.59	481.49	493.57
2016	1100015	3	11024666	31	28	539.91	502.43	526.19
2016	1100015	2	11024682	109	109	541.36	552.07	581.14
2016	1100015	3	11024828	31	38	470.79	464.24	478.05
2016	1100015	3	11024917	15	15	543.34	526.39	535.42

escola

Tentativa incoerente (e intuitiva) de join



- Fazer o join pela possível chave
 - id_municipio e ano

```
SELECT count(*) FROM `enap-331414.enapdatasets.pibpercapita` pp
JOIN `basedosdados.br_inep_ana.escola` ana
ON ana.ano = pp.ano and ana.id_municipio = pp.id_municipio
WHERE pp.ano = 2016
```


- Retorna 47 mil registros
 - Os dados de população e PIB foram replicados para cada escola
 - ✓ Erro comum: join de tabelas incompatíveis
- Quais colunas compõem a chave primária da tabela escola?
 - ano, id_escola
 - ✓ Como chegar numa chave composta por id_municipio e ano ?

Agregar os dados da tabela escola (ANA)

- Não dispensar a dimensão rede
 - Requisito de negócio: suponha que você queira analisar as diferenças entre as redes
 - ✓ Federal, Estadual, Municipal e Privada

ano	id_municipio	rede	id_escola	numero_validos_lp	numero_validos_mt	media_lp_escrita	media_lp_leitura	media_mt
2016	1100015	3	11024372	16	15	475.59	481.49	493.57
2016	1100015	3	11024666	31	28	539.91	502.43	526.19
2016	1100015	2	11024682	109	109	541.36	552.07	581.14
2016	1100015	3	11024828	31	38	470.79	464.24	478.05
2016	1100015	3	11024917	15	15	543.34	526.39	535.42

escola



```
SELECT ano, id_municipio, rede,  
avg(media_lp_escrita) as m_lp_escrita, avg(media_lp_leitura) as m_lp_leitura, avg(media_mt) as m_mat  
FROM `basedosdados.br_inep_ana.escola`  
where ano = 2016  
group by ano, id_municipio, rede
```


Avaliação Nacional da Alfabetização (ANA)

- Objetivo

- adicionar estas informações ao modelo de dados da semana anterior
 - ✓ Aquele com dados de população e PIB
 - Como fazer?

ano	sigla_uf	id_municipio	populacao	nome_municipio	pib
2002	RO	1100023	78039.0	Ariquemes	449592816.0
2003	RO	1100023	79680.0	Ariquemes	539636214.0
2004	RO	1100023	86901.0	Ariquemes	657193231.0
2005	RO	1100023	85031.0	Ariquemes	749021187.0
2006	RO	1100023	86924.0	Ariquemes	790696634.0

pibpercapita

ano	id_municipio	rede	m_lp_escrita	m_lp_leitura	m_mat
2016	1100015	3	501.79111111111111	493.05	495.79444444444444
2016	1100015	2	536.02	542.51	558.505
2016	1100023	2	514.23250000000001	516.92500000000001	518.15249999999999
2016	1100023	3	491.7919047619048	497.53428571428566	497.94523809523804
2016	1100031	3	506.23666666666666	501.76	494.23333333333335
2016	1100049	3	511.42117647058825	514.3376470588236	517.5670588235295

agregação de escola

Rede

- 1 = Federal
- 2 = Estadual
- 3 = Municipal
- 4 = Privada

Pivot Table no BigQuery

- Acesso somente com usuário autenticado (Conta Google)
 - <https://console.cloud.google.com/bigquery?project=enap-331414>
 - ✓ [Documentação do BigQuery](#)

product	sales	quarter
Kale	51	Q1
Kale	23	Q2
Kale	45	Q3
Kale	3	Q4
Apple	77	Q1
Apple	0	Q2
Apple	25	Q3
Apple	2	Q4



product	Q1	Q2	Q3	Q4
Apple	77	0	25	2
Kale	51	23	45	3

```
SELECT * FROM
  (SELECT * FROM Produce)
 PIVOT(SUM(sales) FOR quarter IN ('Q1', 'Q2', 'Q3', 'Q4'))
```

Pivotear a agragação da tabela escola (ANA)

ano	id_municipio	rede	m_lp_escrita	m_lp_leitura	m_mat
2016	1100015	3	501.79111111111111	493.05	495.79444444444444
2016	1100015	2	536.02	542.51	558.505
2016	1100023	2	514.23250000000001	516.92500000000001	518.15249999999999
2016	1100023	3	491.7919047619048	497.53428571428566	497.94523809523804
2016	1100031	3	506.23666666666666	501.76	494.23333333333335
2016	1100049	3	511.42117647058825	514.3376470588236	517.5670588235295

agregação de escola

Rede

- 1 = Federal
- 2 = Estadual
- 3 = Municipal
- 4 = Privada

SELECT * FROM


```
(
    SELECT ano, id_municipio, rede,
    avg(media_lp_escrita) as m_lp_escrita, avg(media_lp_leitura) as m_lp_leitura, avg(media_mt) as m_mat
    FROM `basedosdados.br_inep_ana.escola`
    group by ano, id_municipio, rede
)
PIVOT(SUM(m_lp_escrita) as mescrita, SUM(m_lp_leitura) as mleitura, SUM(m_mat) as mmatematica FOR rede IN ('1','2','3','4'))
```

ano	id_municipio	mescrita_1	mleitura_1	mmatematica_1	mescrita_2	mleitura_2	mmatematica_2
2016	1100015	null	null	null	536.02	542.51	558.505
2016	1100023	null	null	null	514.23250000000001	516.92500000000001	518.15249999999999
2016	1100031	null	null	null	null	null	null
2016	1100049	null	null	null	521.80125000000001	535.95499999999999	543.1575
2016	1100056	null	null	null	547.7	574.48	571.78
2016	1100064	null	null	null	518.0825	534.585	532.3825

agregação de escola
pivotada

Join com a tabela pibpercapita

```
SELECT pp.nome_municipio, pp.populacao, pp.pib, escola.* FROM
(
  SELECT ano, id_municipio, rede,
  avg(media_lp_escrita) as m_lp_escrita, avg(media_lp_leitura) as m_lp_leitura, avg(media_mt) as m_mat
  FROM `basedosdados.br_inep_ana.escola`
  group by ano, id_municipio, rede
)
PIVOT(SUM(m_lp_escrita) as mescrita, SUM(m_lp_leitura) as mleitura,
      SUM(m_mat) as mmatematica FOR rede IN ('1','2','3','4')) as escola
JOIN `enap-331414.enapdatasets2.pibpercapita` pp on pp.ano = escola.ano and pp.id_municipio = escola.id_municipio
LIMIT 10
```



populacao	pib	ano	id_municipio	mescrita_1	mleitura_1	mmatematica_1	mescrita_2
801718.0	1.8740050091E10	2016	2507507	534.47	580.08	594.28	445.147894736842
2491109.0	8.7248917712E10	2014	3106200	576.9	614.22	603.31	529.5405147058822
2513451.0	8.8397460694E10	2016	3106200	571.37	584.38	579.59	536.7118796992484
1472482.0	6.3989576204E10	2014	4314902	559.23	596.73	583.33	492.971933962264
1481019.0	7.2734881188E10	2016	4314902	546.11	549.12	554.17	491.2806372549021
461524.0	1.6915925686E10	2014	4205407	579.16	604.56	604.59	523.2264285714284
477798.0	1.8660875642E10	2016	4205407	579.82	585.48	584.18	505.496

resultado do join

Exercício 5.1

- Reutilizando a Query do exercício 3.1
 - Aquela que faz o join da população com o PIB dos municípios
- Incremente sua query fazendo mais um join
 - com os dados do ideb dos municípios presentes no repositório
 - ✓ basedosdados:br_inep_ideb.municipio
 - Do BigQuery
 - Adicione ao seu modelo de dados a informação da `nota_saeb_media_padronizada`
 - ✓ das redes Municipal, Estadual, Federal e Pública
- [Link](#) para submissão dos exercícios 5.1 e 5.2
- Deadline de entrega: 22/11, 14h30

Exercício 5.2

- Reutilizando o dataframe do exercício 3.1
 - Aquele que contém os dados da população e do pib dos municípios
- Incremente seu dataframe fazendo join/merge, com o pandas,
 - com os dados do ideb dos municípios presentes no repositório
 - ✓ basedosdados:br_inep_ideb.municipio
 - Do BigQuery
 - Adicione ao seu modelo de dados a informação da `nota_saeb_media_padronizada`
 - ✓ das redes Municipal, Estadual, Federal e Pública
 - Faça, no pandas, somente o merge dos dataframes
- Elenque critérios a serem considerados para escolher entre
 - A solução do exercício 5.1 ou a solução do exercício 5.2

Se entender, já está falando a língua dos nerds

