DATA SCIENCE
PPGIA/PUCPR

Prof. Jean Paul Barddal

SCIENTIA, VITA ET FIDES
PUCPR
GRUPO MARISTA

NOTES ABOUT PRACTICAL TEST

Practical test

- You have now selected a dataset
- Use the template provided to perform your practical test
- You MUST deliver your practical test by the end of our last lecture via email (jean.barddal@ppgia.pucpr.br)
- Remember: you will also have a theoretical test to be made on our last lecture (April 19th)

EXPLORATORY DATA ANALYSIS VERSUS EXPLANATORY DATA ANALYSIS

Exploratory analysis vs. Explanatory analysis

- Exploratory
  - Analysis conducted when we need to understand the data
  - Questions are made and we answer them using statistics or visualizations
  - Visualizations are not perfect
- Explanatory
  - Aims at “polishing” the results of the explatory analysis
  - Highlights the insights obtained
  - Is often coupled with a story or demand

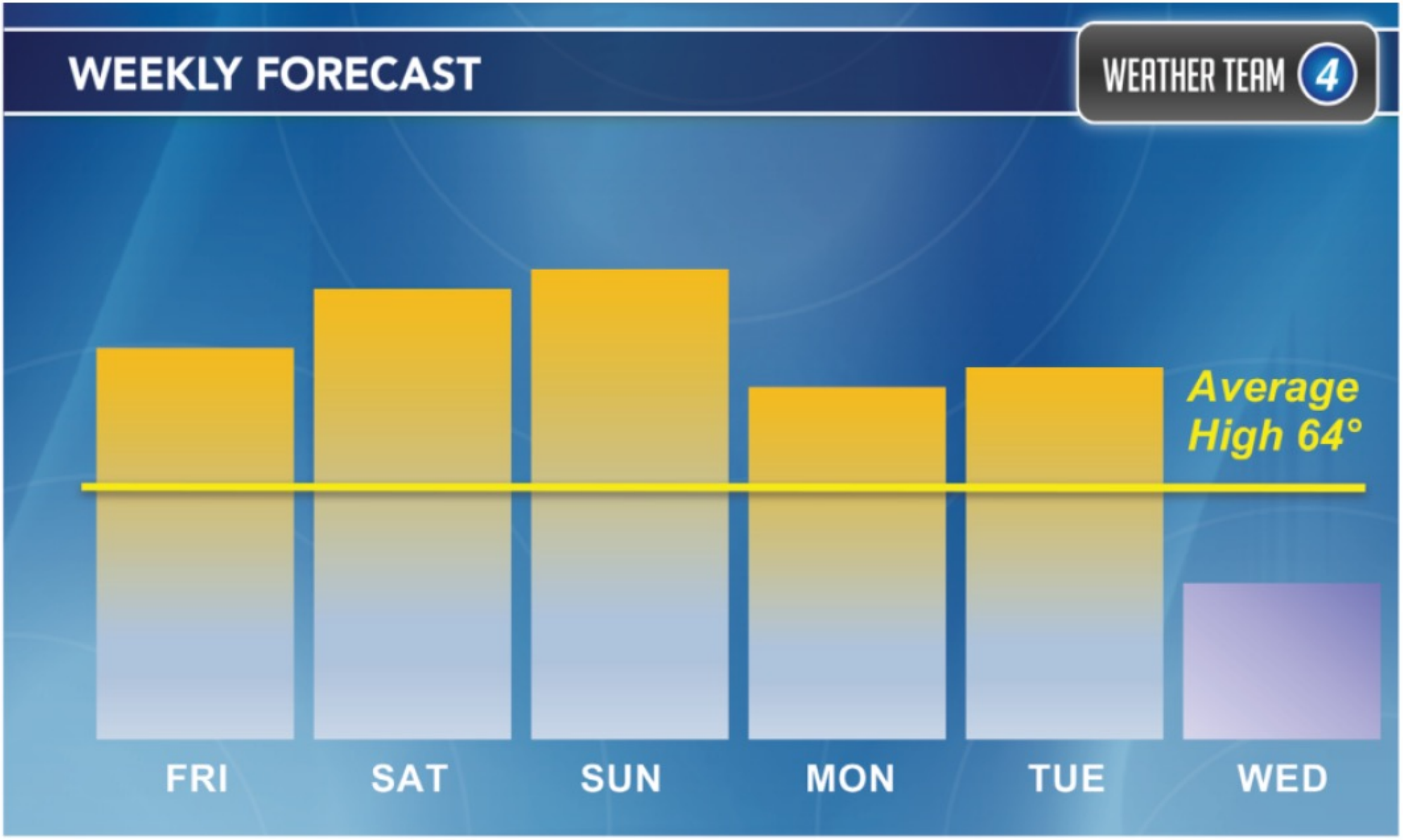
Steps

- Data extraction
- Data cleansing
- Exploratory analysis
- Data analysis
- **Sharing**

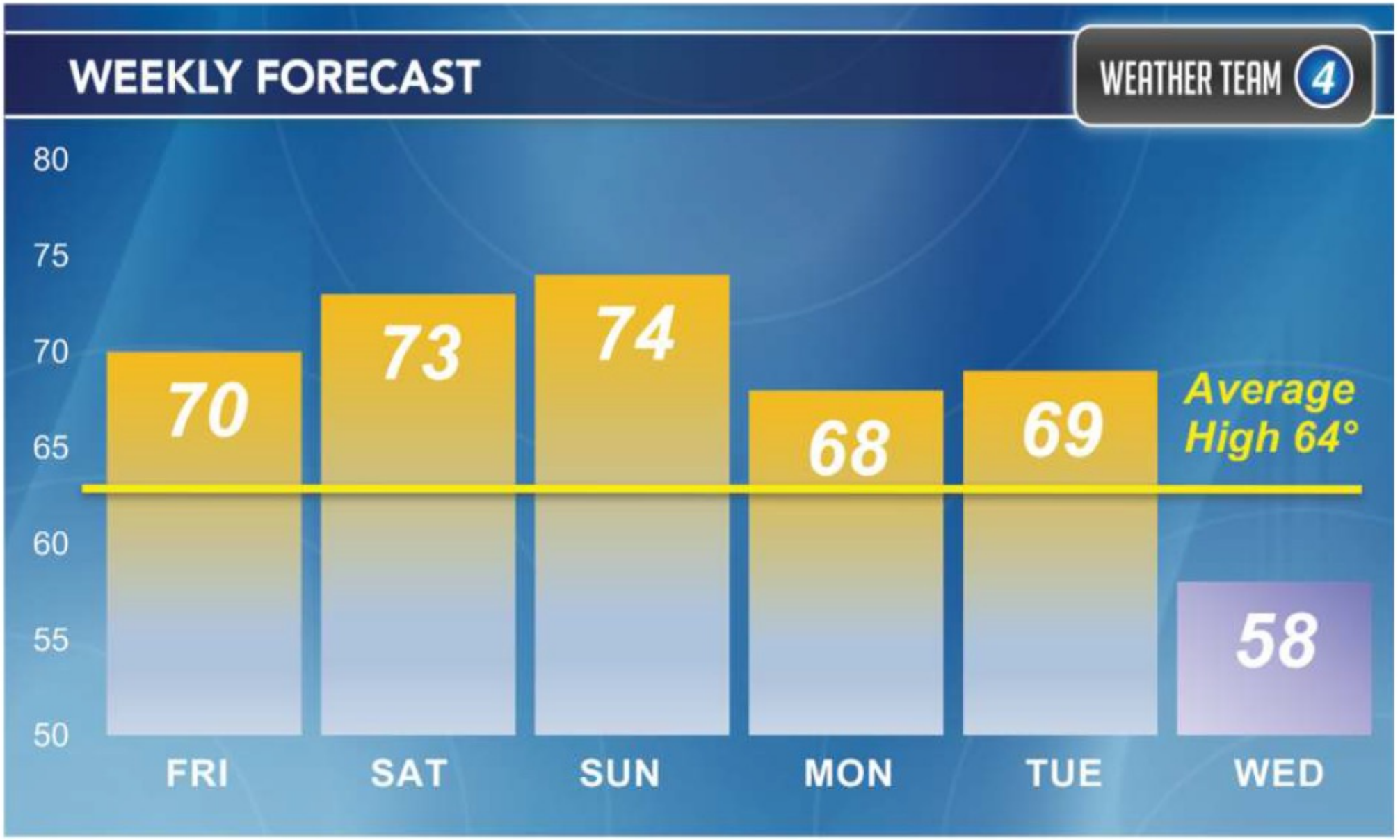
I have a dataset and I need to present it to someone else.

How can I do so, **effectively**?

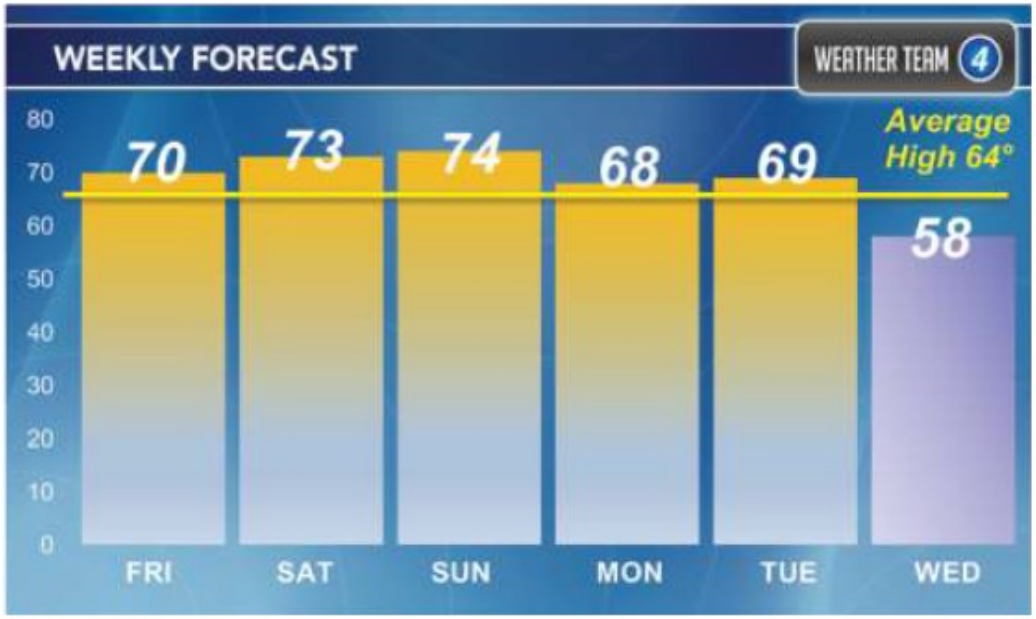
Analyze the following image.
What can you infer about it?


WEEKLY FORECAST
WEATHER TEAM 4
Average
High 64°
FRI
SAT
SUN
MON
TUE
WED

What temperature would you estimate for Sunday?


WEEKLY FORECAST
WEATHER TEAM 4
80
75
70
65
60
55
50
70
73
74
68
69
58
Average
High 64°
FRI
SAT
SUN
MON
TUE
WED


WEEKLY FORECAST
WEATHER TEAM 4
80
75
70
65
60
55
50
70
73
74
68
69
58
Average
High 64°
FRI
SAT
SUN
MON
TUE
WED


WEEKLY FORECAST
WEATHER TEAM 4
80
70
60
50
40
30
20
10
0
70
73
74
68
69
58
Average
High 64°
FRI
SAT
SUN
MON
TUE
WED

EFFECTIVE DATA VIZUALIZATION

Effective data visualization

- Visualizations are means to communicate, and thus, we must ensure that the reader acknowledges the same information we intended to divulge
- Suggestion: triple-check the checklist that comes next
- We will work on this topic following a "*reductio ad absurdium*" approach in the sense that we will check what should **NOT** be done

Checklist

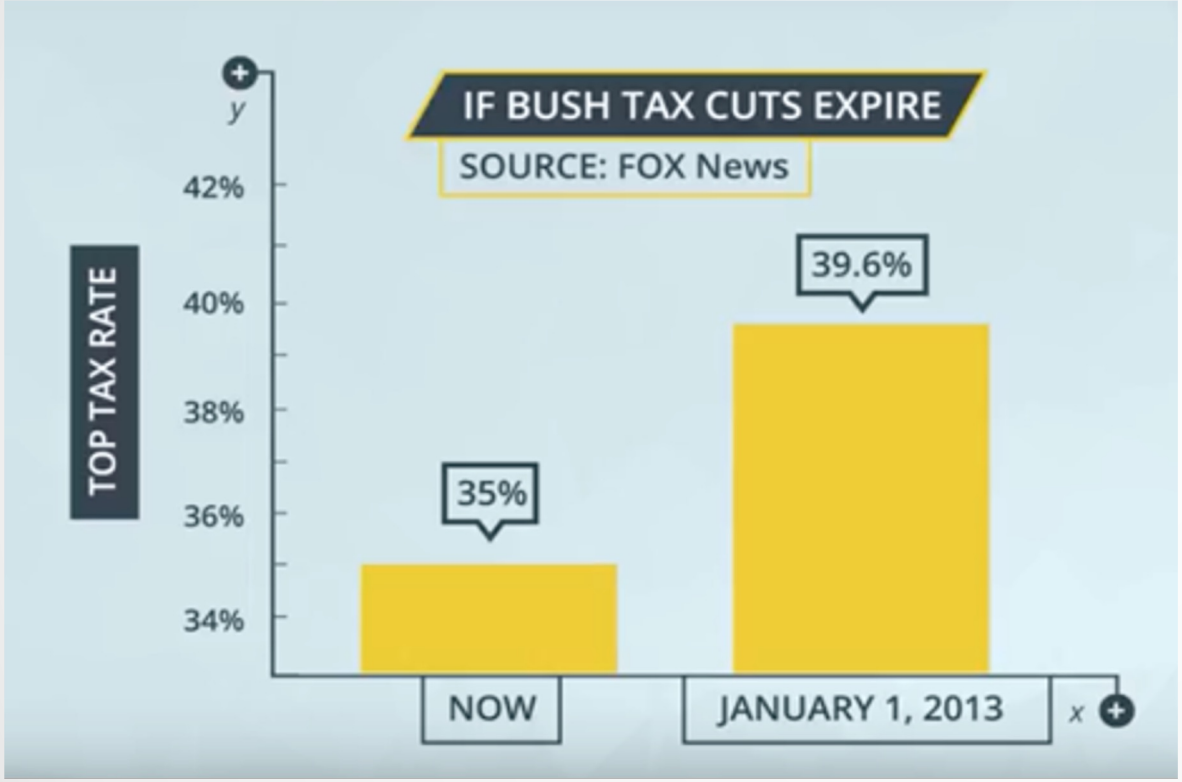
- Title
- Axes labels
- Axes units
- Legend
- Scale
- Order
- Colors
- Text size
- Chart junk

POOR DATA VISUALIZATION

What deems a visualization poor?

- A visualization is poor if our message is unclear
- This means we should avoid:
  - Ambiguity
  - Lack of information
  - Omission
  - Distraction

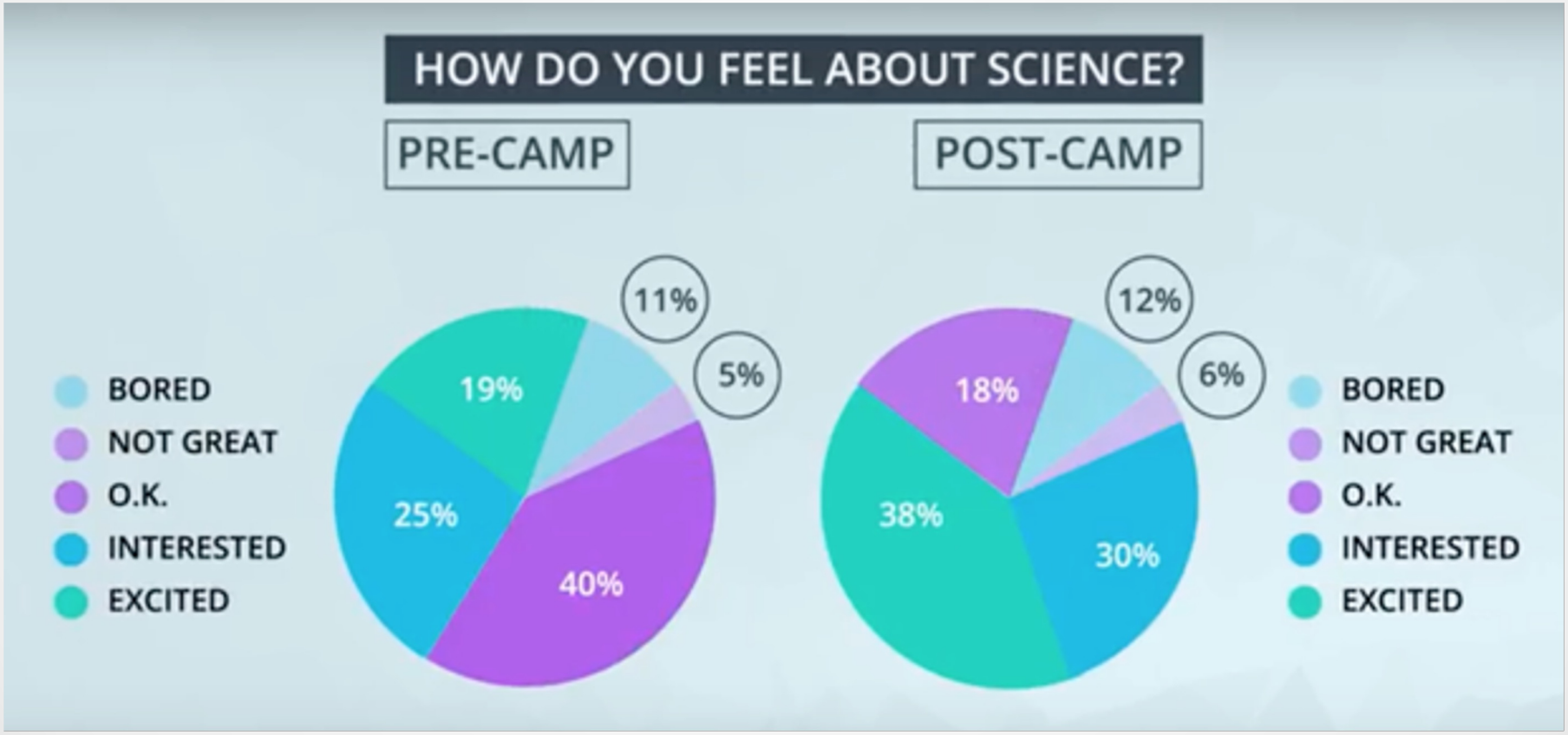
Example



- The difference between the bars is somewhat small, but the scale makes us believe the difference is huge

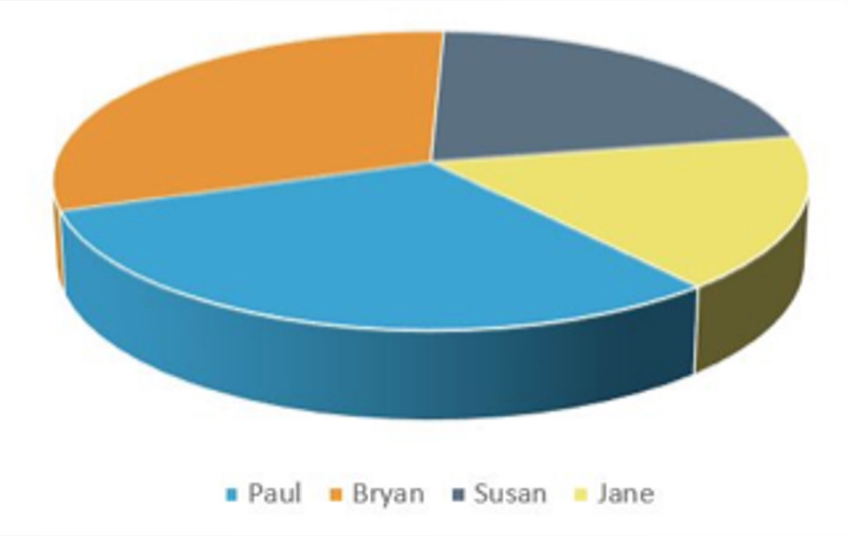
Example

Is it possible to say that the interest in science increased with the camp?

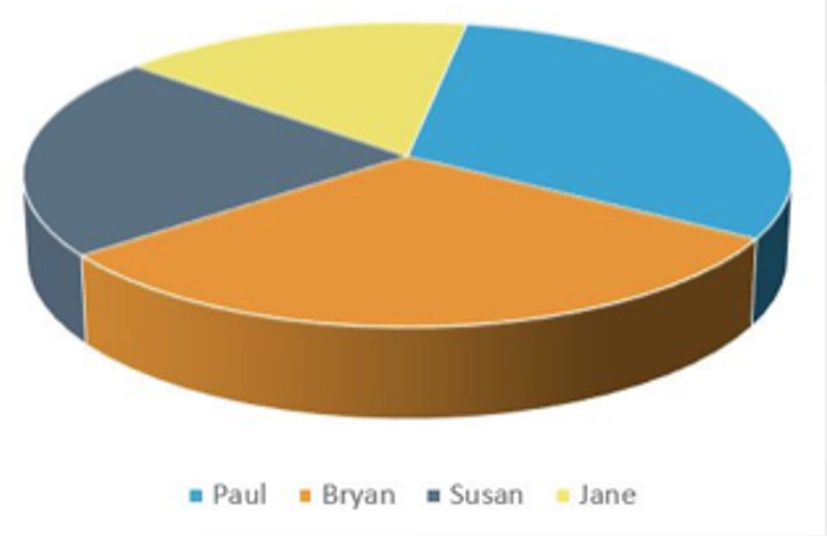


Example

Which of the regions below is larger?



Bryan or Paul?



What about here?

Visual components

Data visualizations are tailored using the following components:

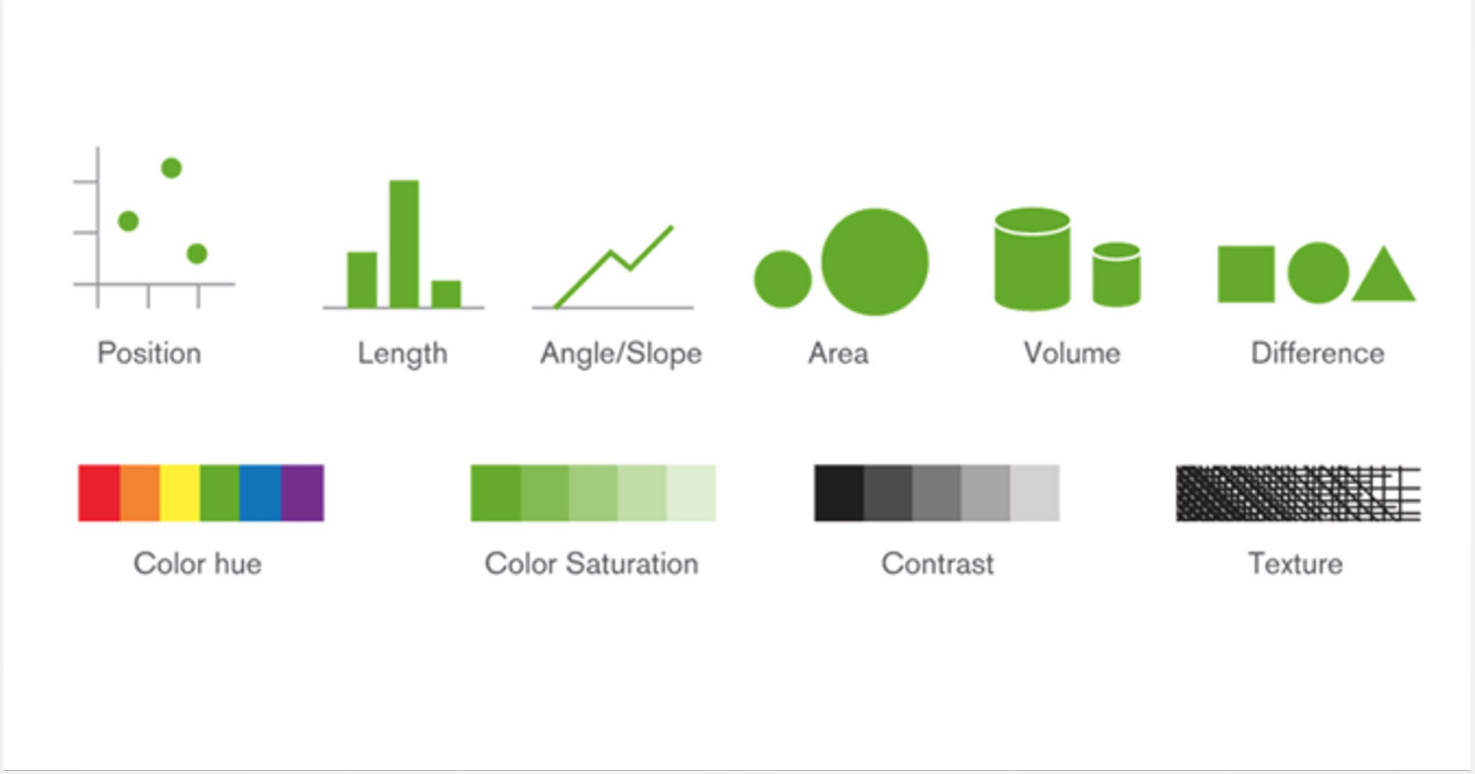


Chart Junk

- When deciding which components we should use, it is a good idea to think about what **NOT** to use
- Tufte created the "data-ink ratio" concept, which is the relation between the ink required to plot the data and the ink used for the rest

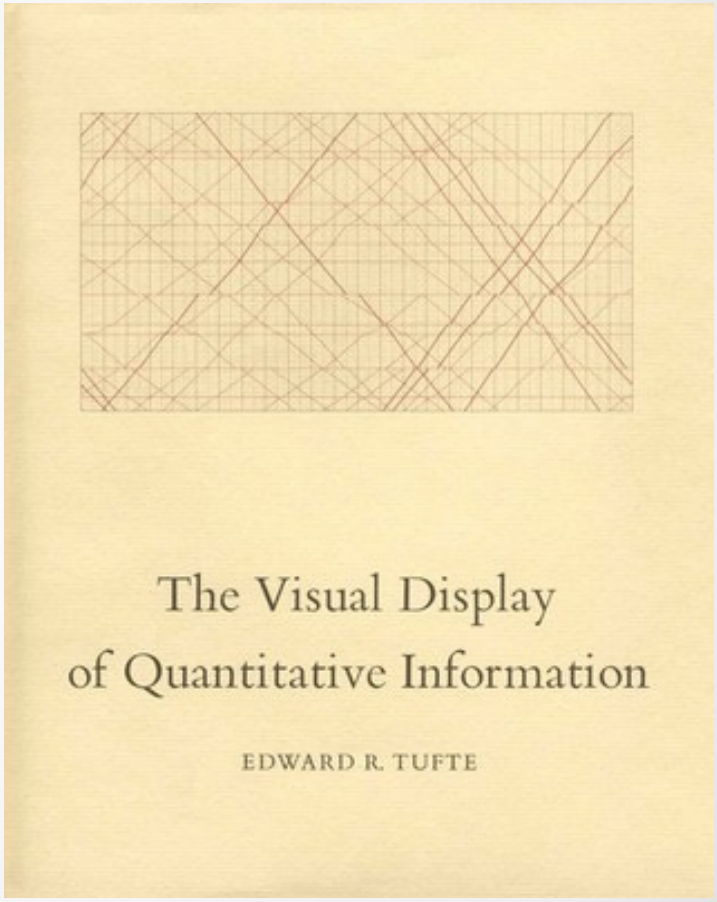


Chart Junk

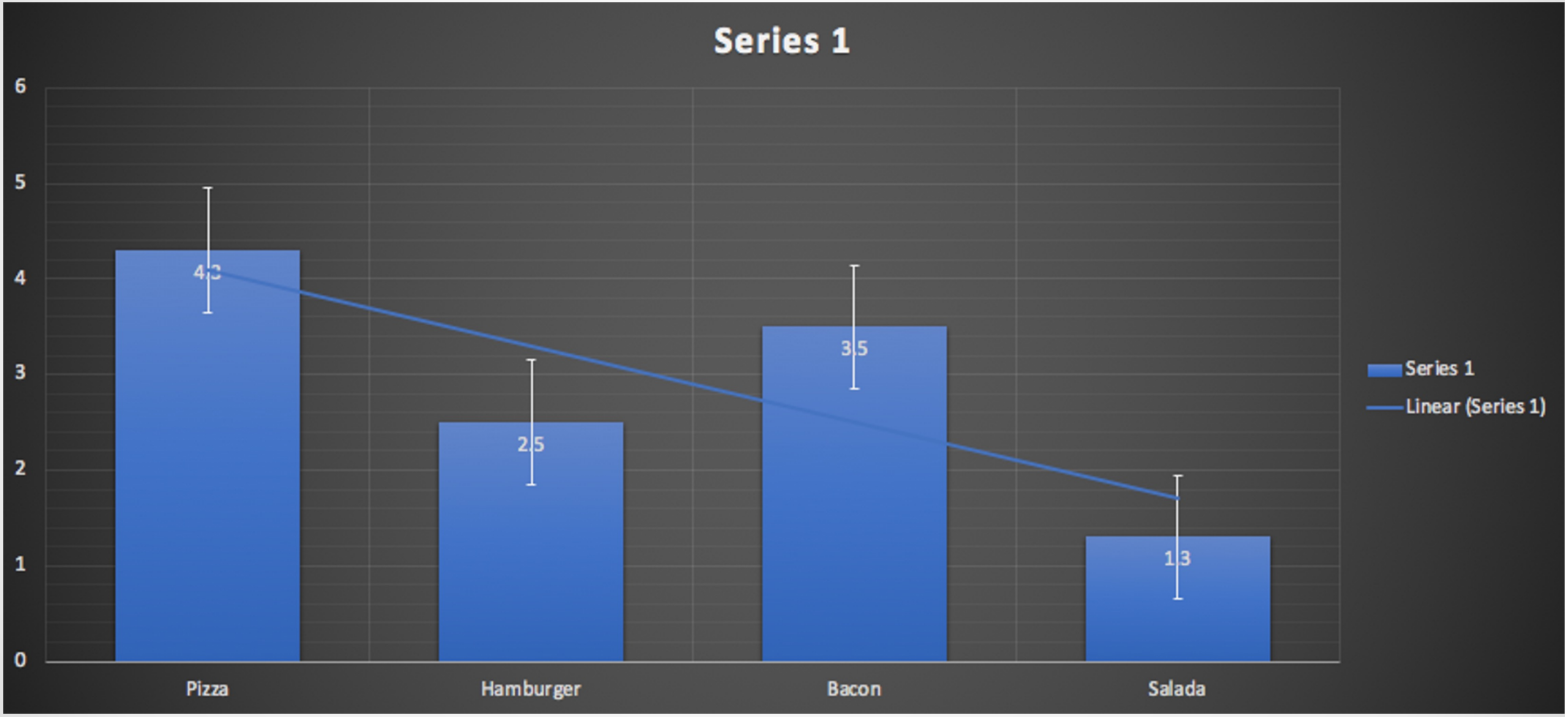

Series 1
6
5
4
3
2
1
0
4.3
2.5
3.5
1.3
Pizza
Hamburger
Bacon
Salada
Series 1
Linear (Series 1)

Chart Junk

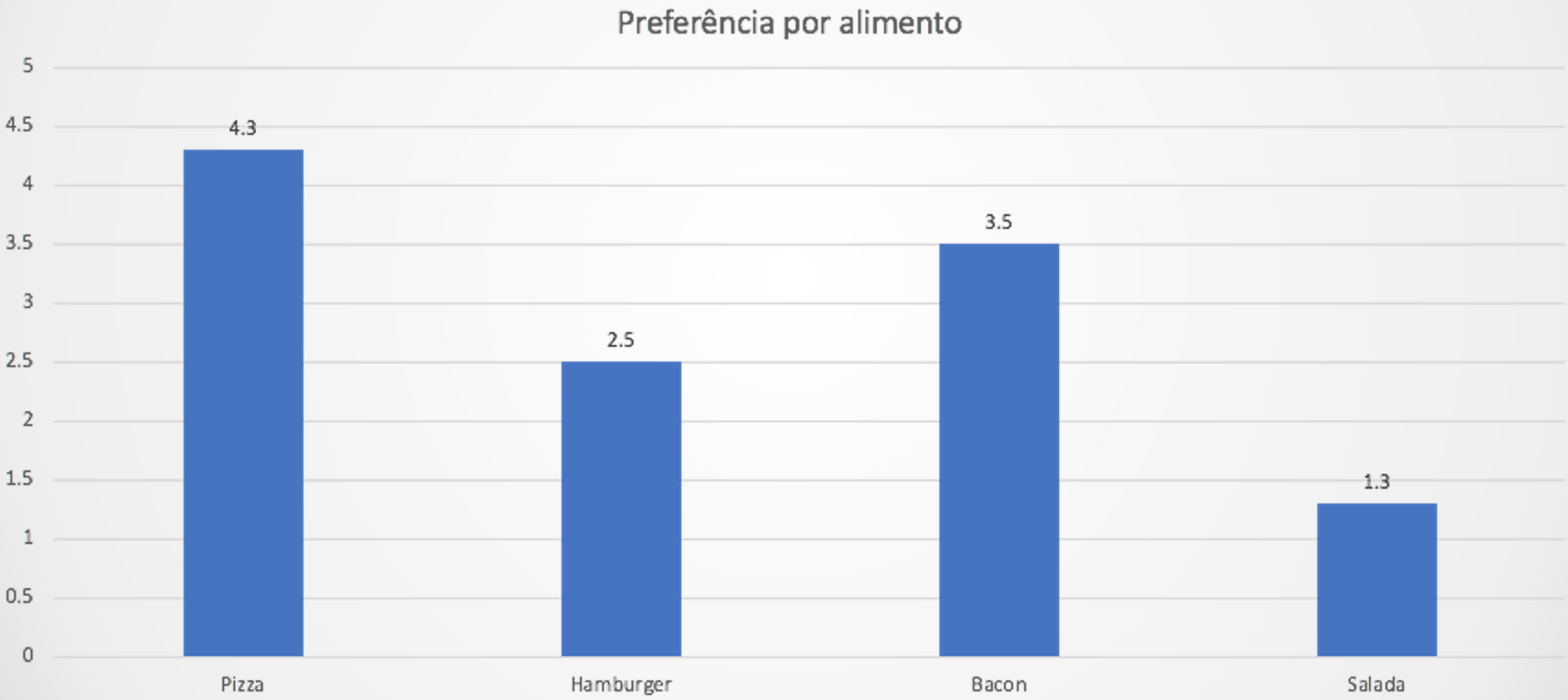

Preferência por alimento
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
4.3
2.5
3.5
1.3
Pizza
Hamburger
Bacon
Salada

Chart Junk

Colors can be used to highlight something!

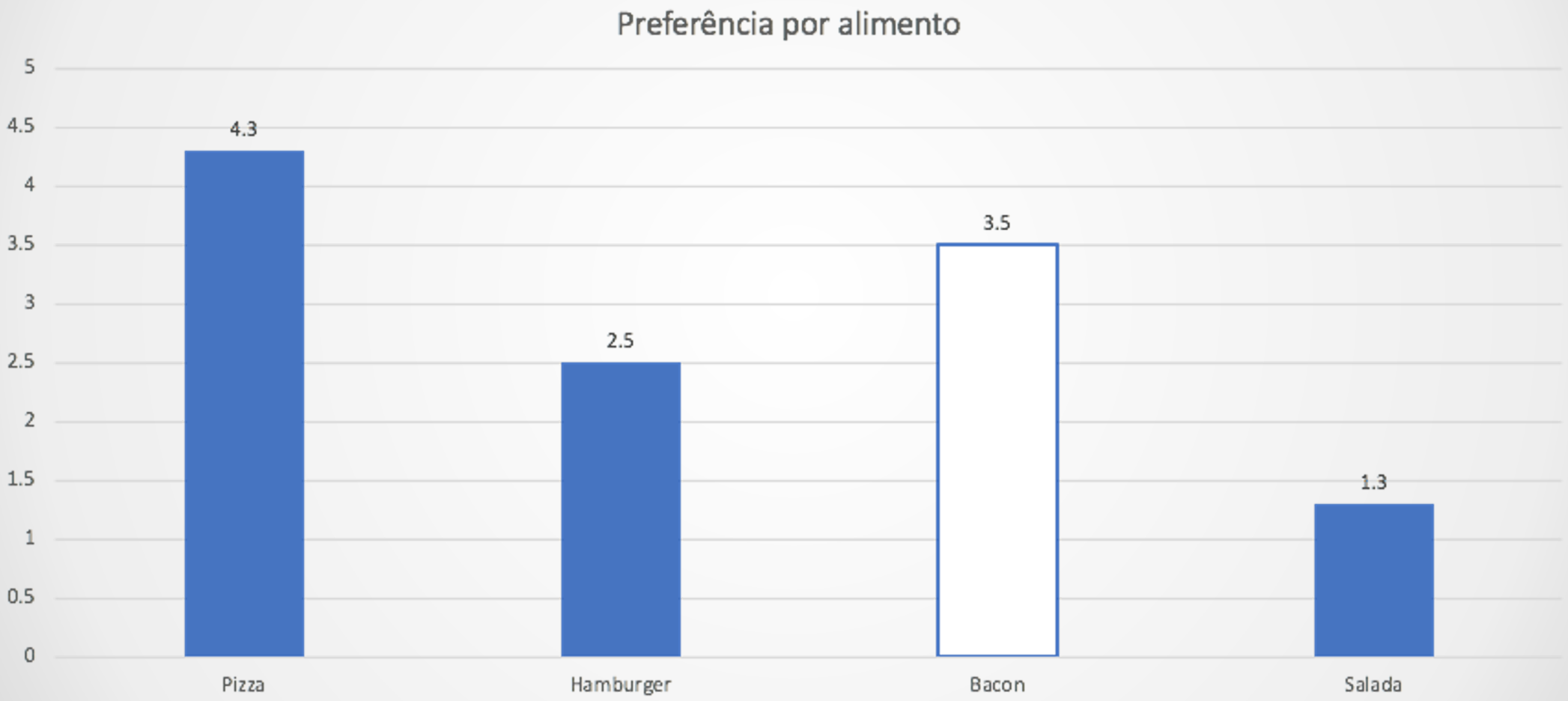
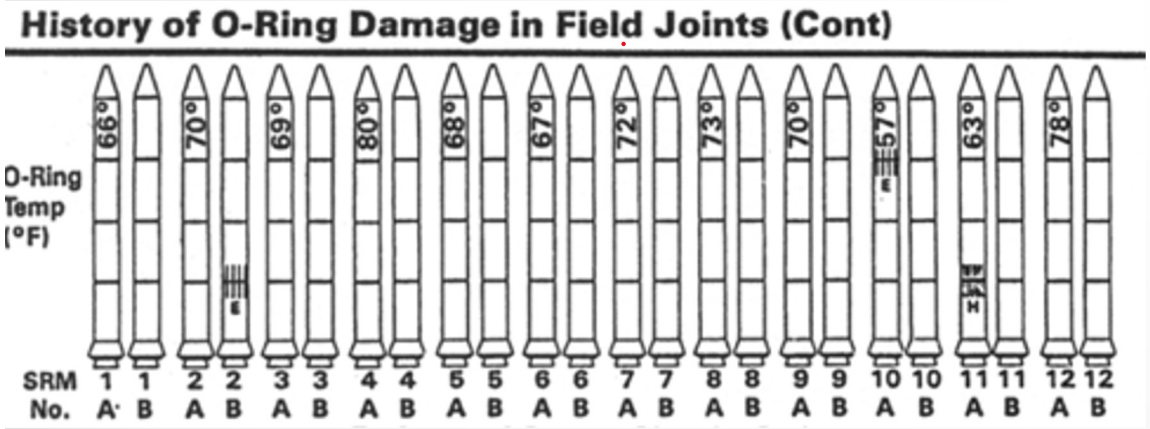
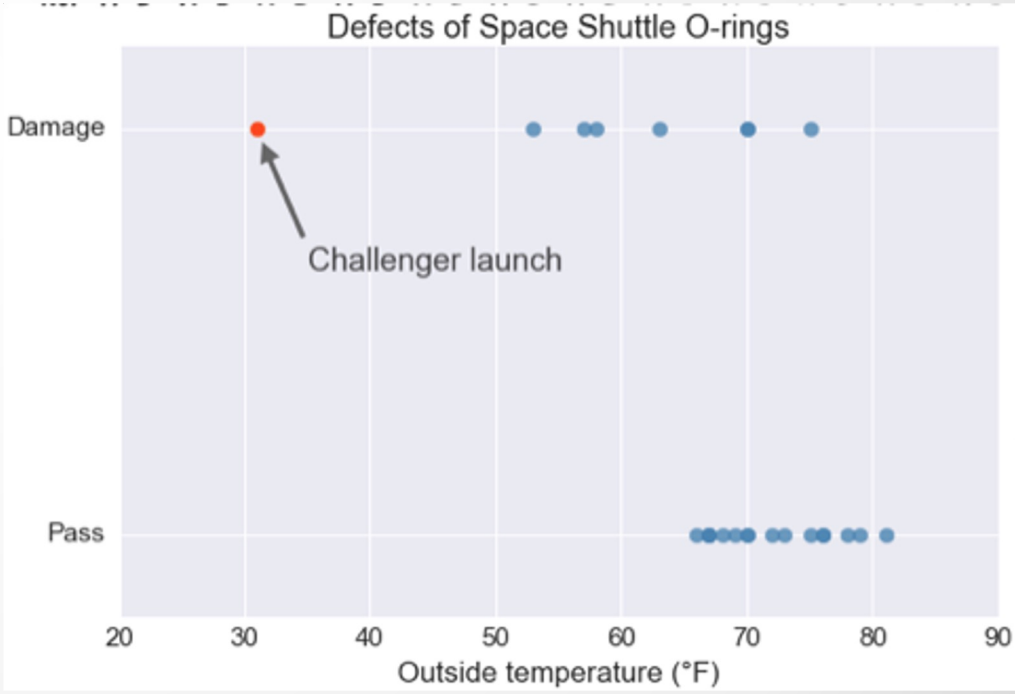


Chart Junk

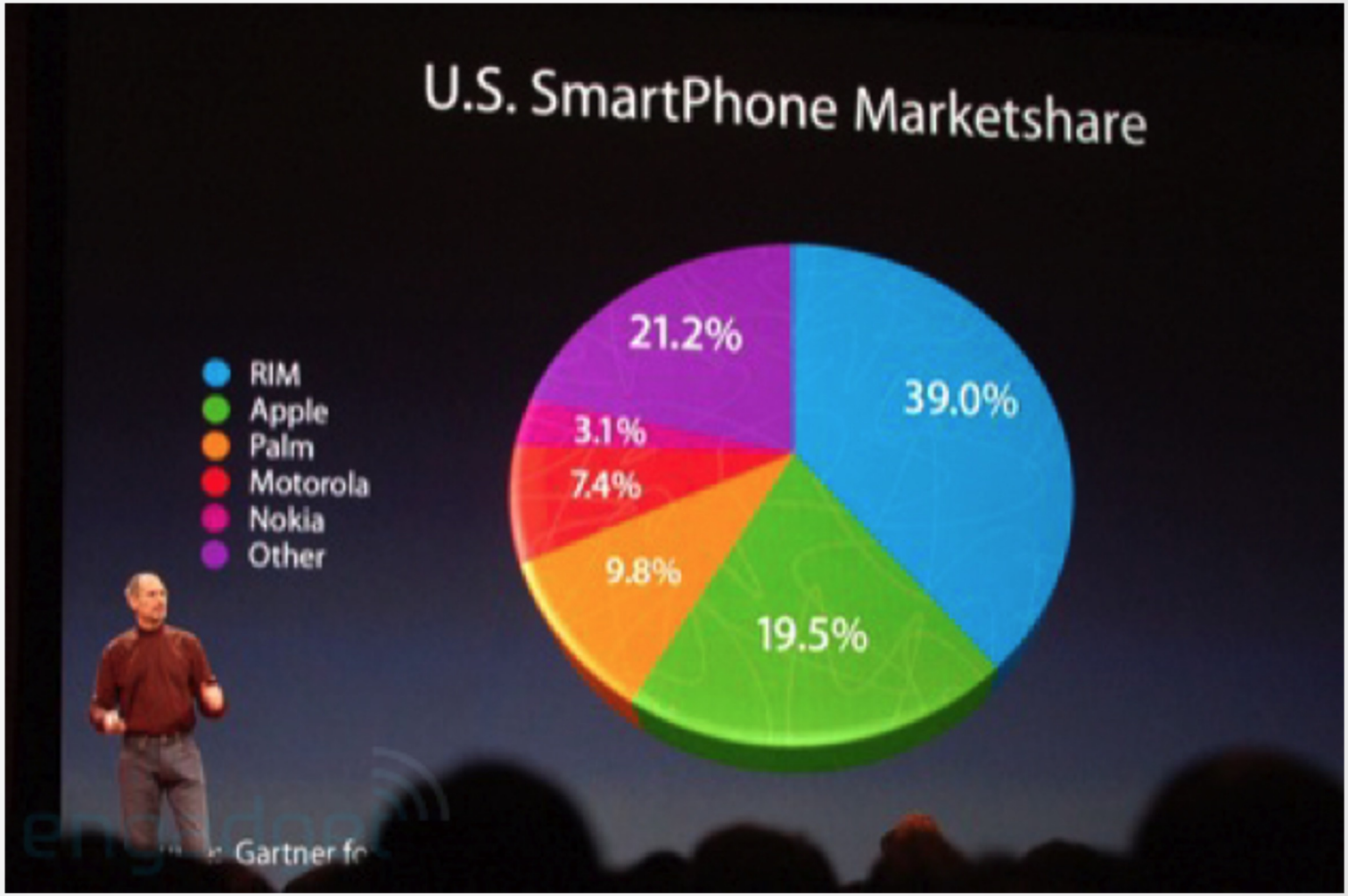
What is your feeling about these plots?



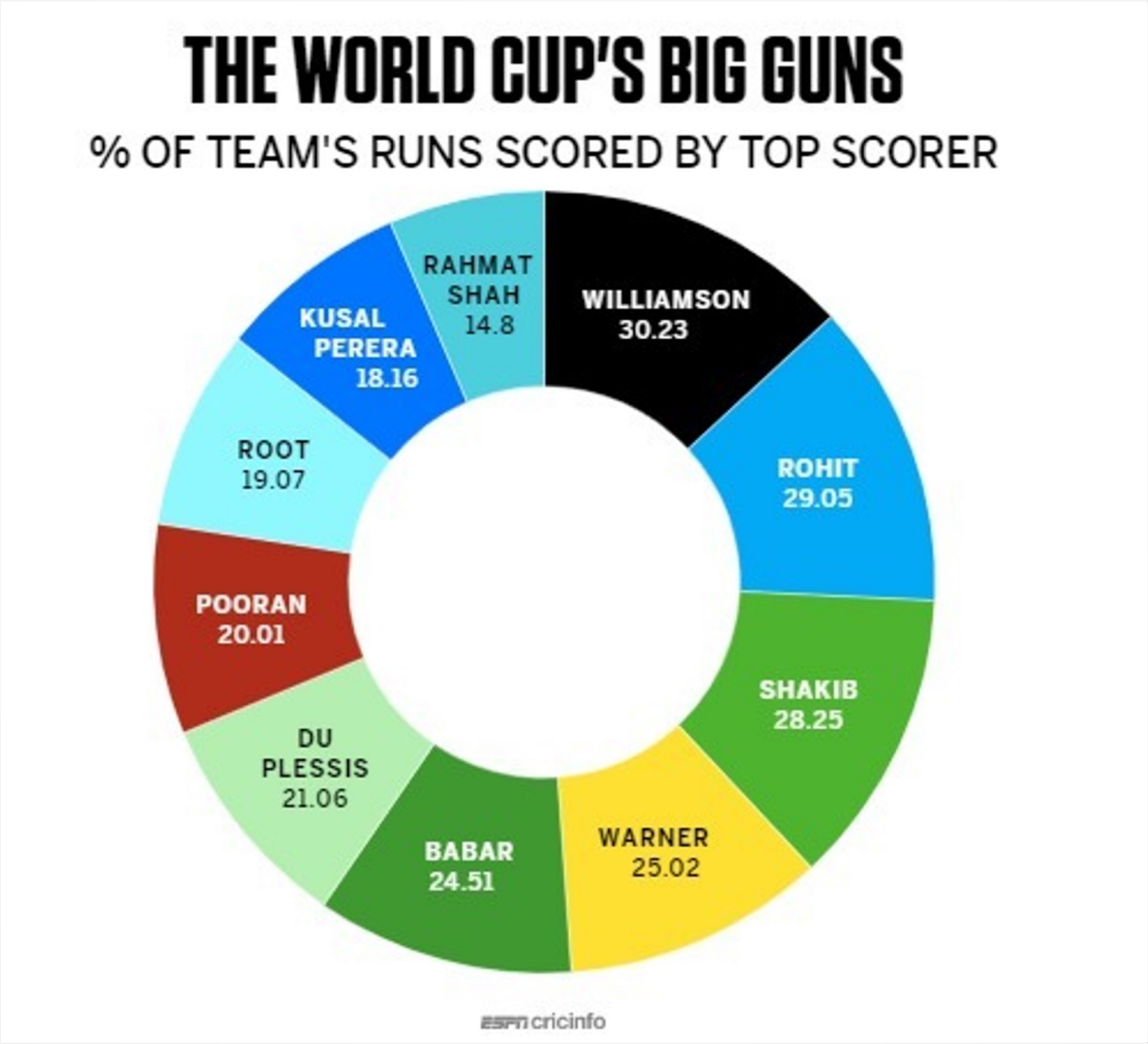


ANALYSIS

What problems do you see here?



And here?


THE WORLD CUP'S BIG GUNS
% OF TEAM'S RUNS SCORED BY TOP SCORER
RAHMAT SHAH 14.8
WILLIAMSON 30.23
KUSAL PERERA 18.16
ROHIT 29.05
ROOT 19.07
POORAN 20.01
SHAKIB 28.25
DU PLESSIS 21.06
BABAR 24.51
WARNER 25.02
ESPN cricinfo

Mortes de Covid-19 no Brasil

75% das vítimas tinham doenças associadas

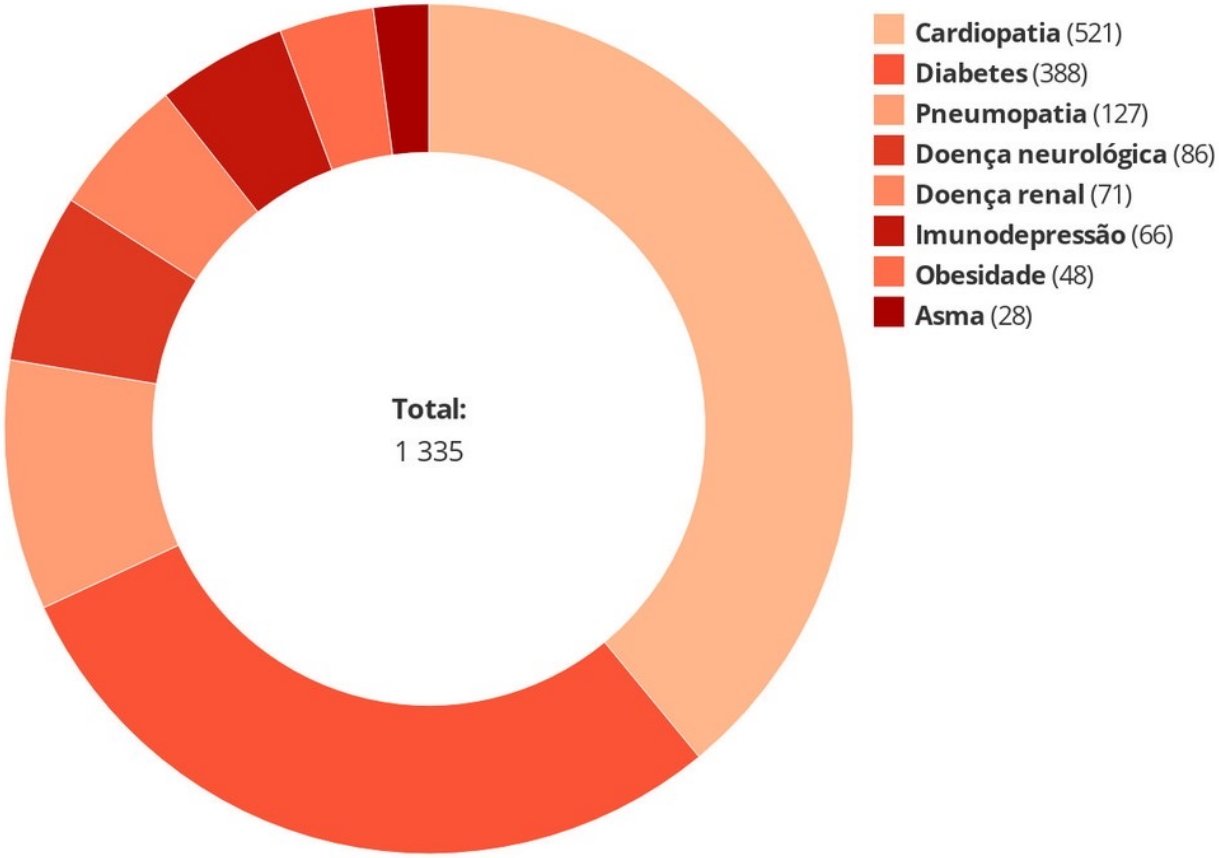
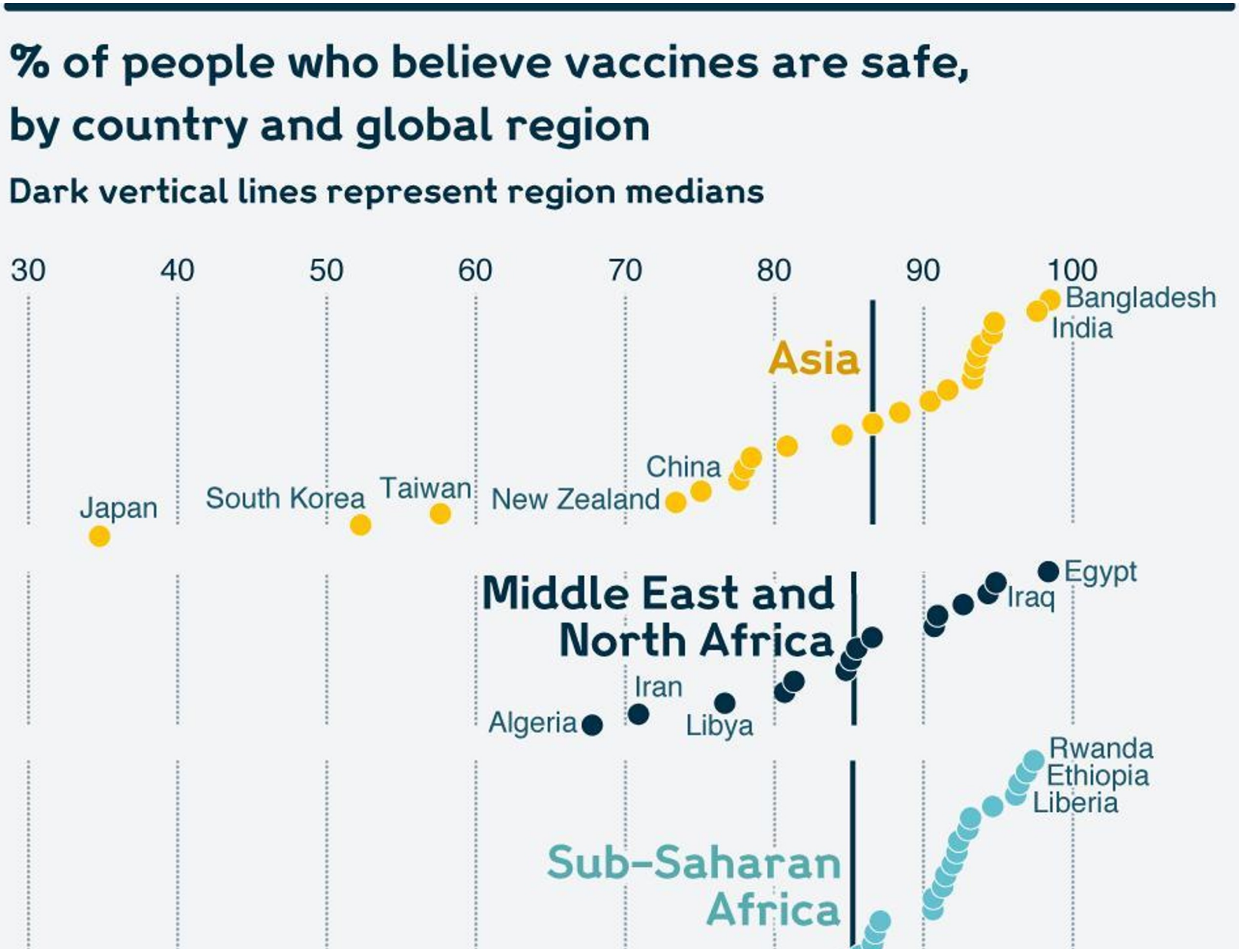
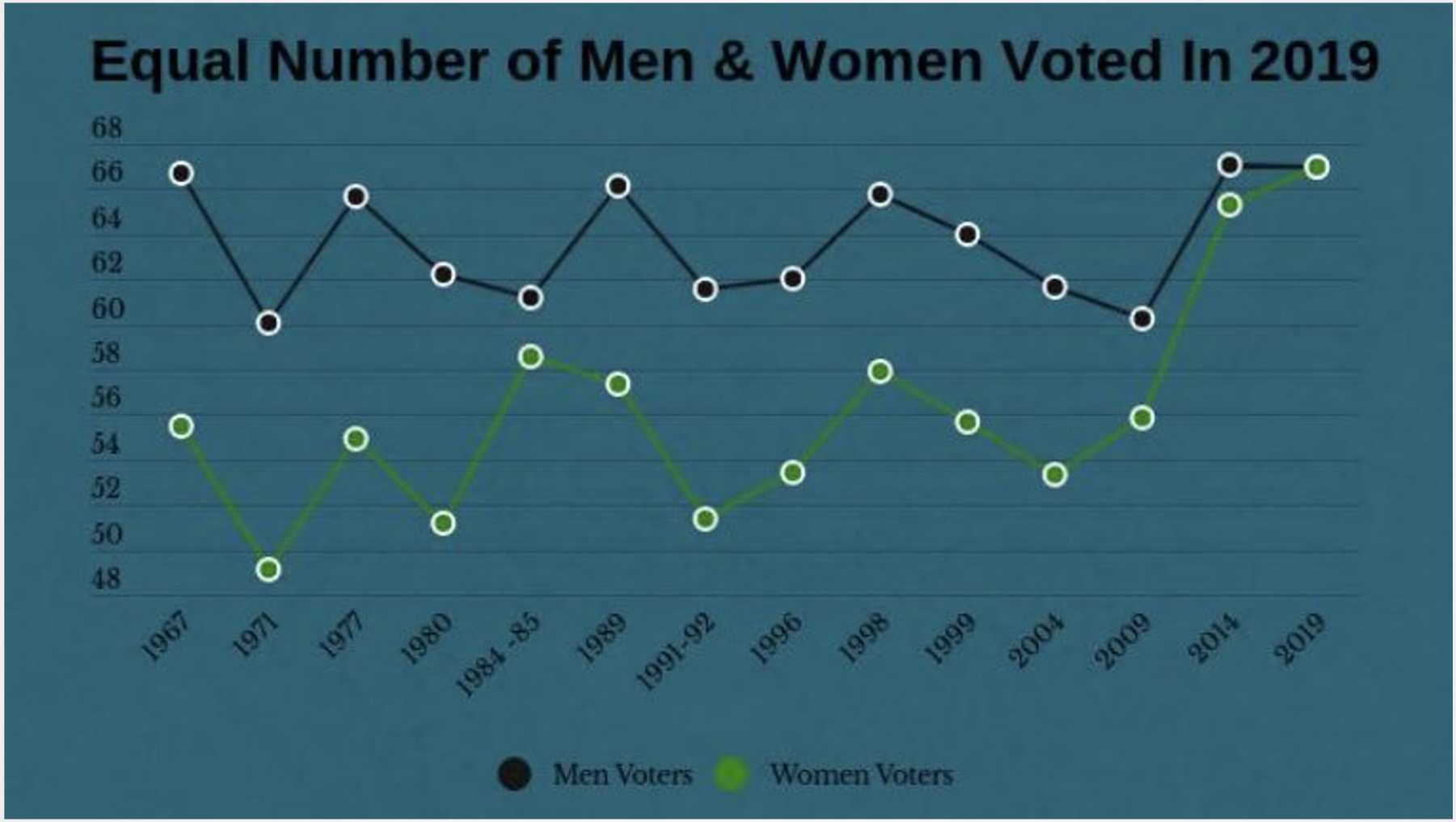
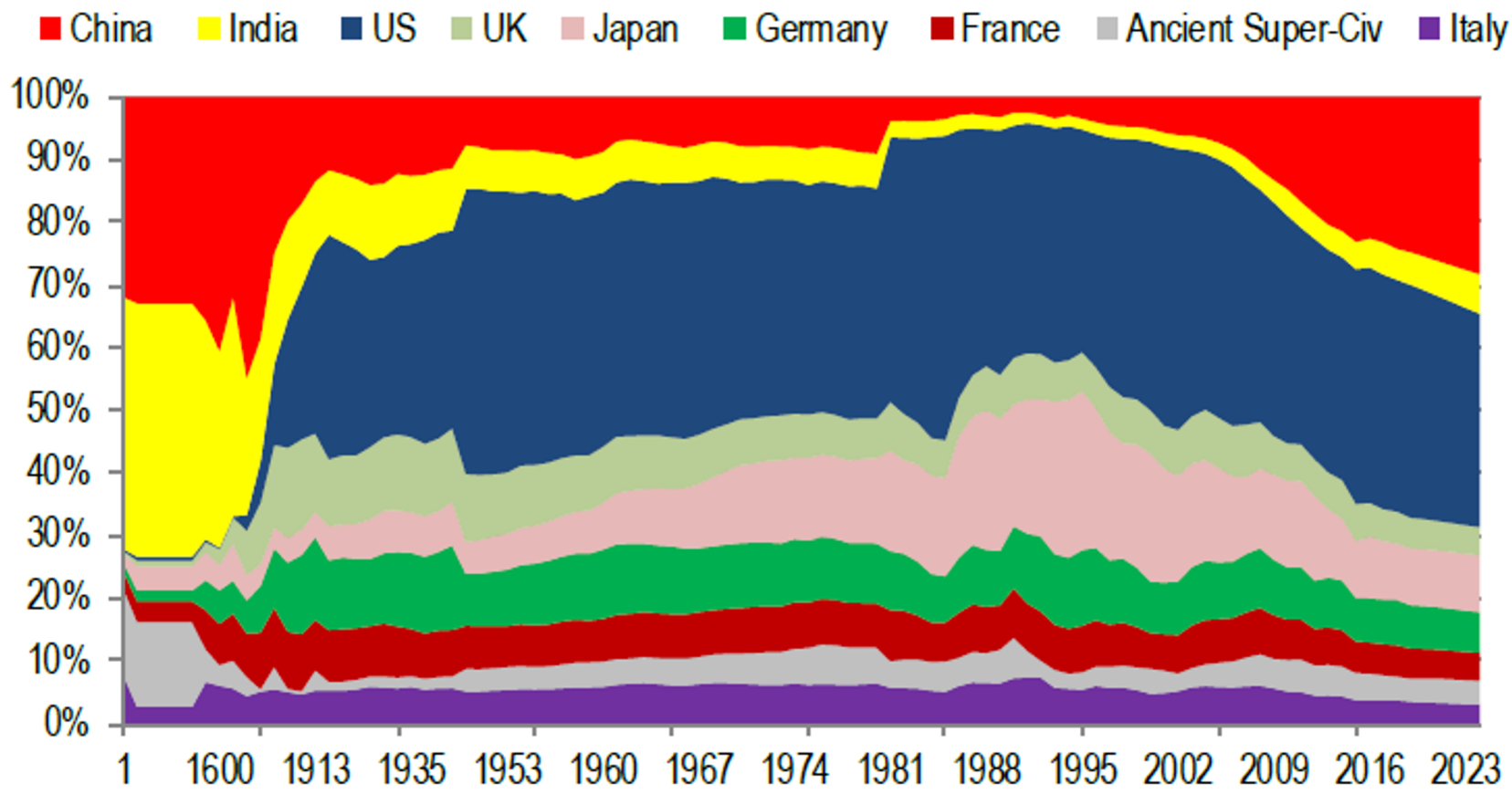


Gráfico: G1 • Fonte: Ministério da Saúde

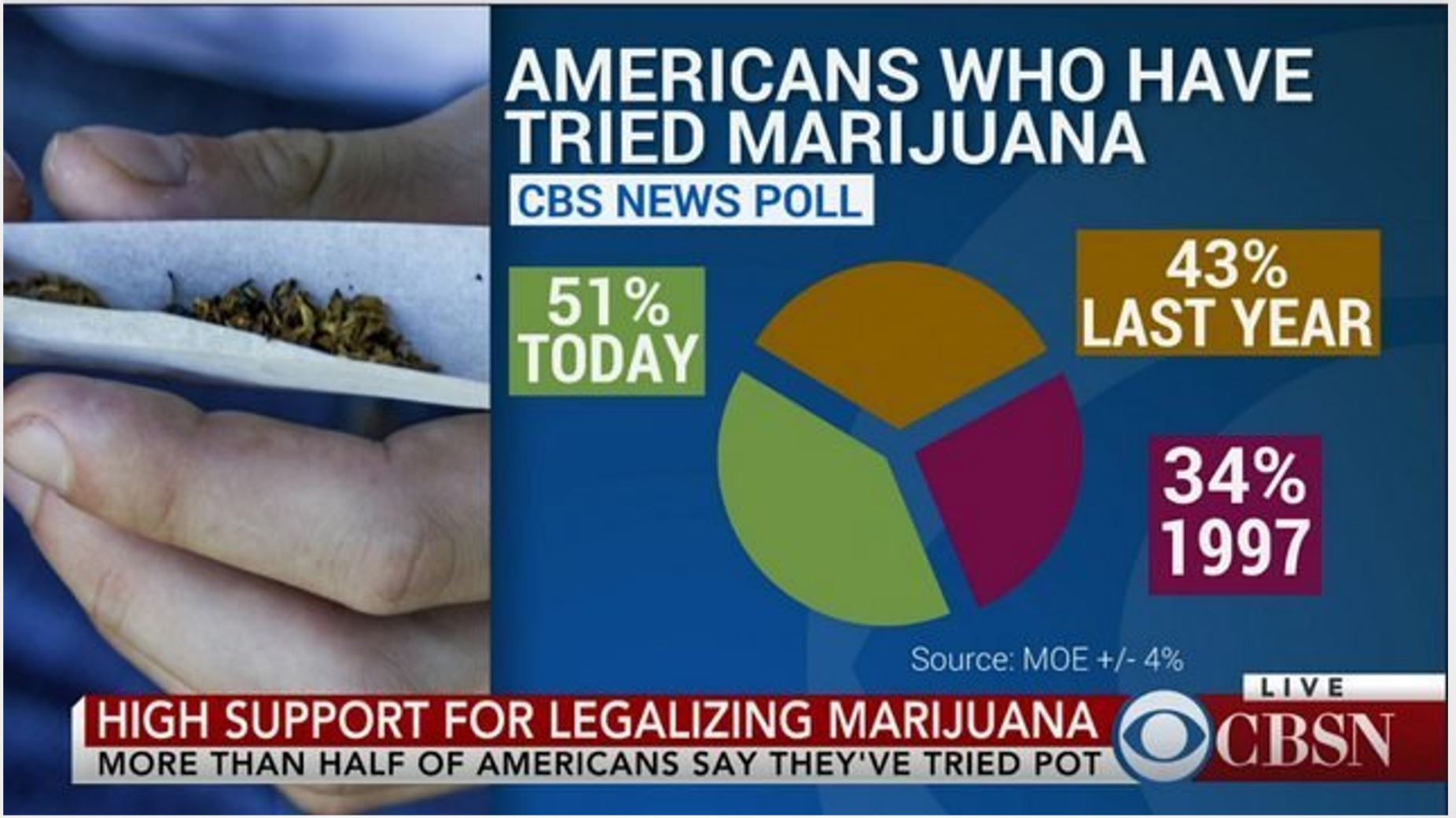

% of people who believe vaccines are safe, by country and global region
Dark vertical lines represent region medians
30
40
50
60
70
80
90
100
Asia
Bangladesh
India
China
New Zealand
Taiwan
South Korea
Japan
Middle East and North Africa
Egypt
Iraq
Iran
Algeria
Libya
Sub-Saharan Africa
Rwanda
Ethiopia
Liberia


Equal Number of Men & Women Voted In 2019
68
66
64
62
60
58
56
54
52
50
48
1967
1971
1977
1980
1984 -85
1989
1991-92
1996
1998
1999
2004
2009
2014
2019
Men Voters
Women Voters

The progression of world GDP – percentage of world GDP by major countries/regions. Are current trade tensions really an attempt to stymie China's impressive rise?



Source: BofA Merrill Lynch Global Research, Angus Maddison, IMF GDP data and estimates between 1980 and 2024. Ancient super-civilizations include Greece, Turkey, Algeria, Iraq, Egypt and Iran.


AMERICANS WHO HAVE TRIED MARIJUANA
CBS NEWS POLL
51% TODAY
43% LAST YEAR
34% 1997
Source: MOE +/- 4%
HIGH SUPPORT FOR LEGALIZING MARIJUANA
MORE THAN HALF OF AMERICANS SAY THEY'VE TRIED POT
LIVE
CBSN

If you’re not satisfied enough...

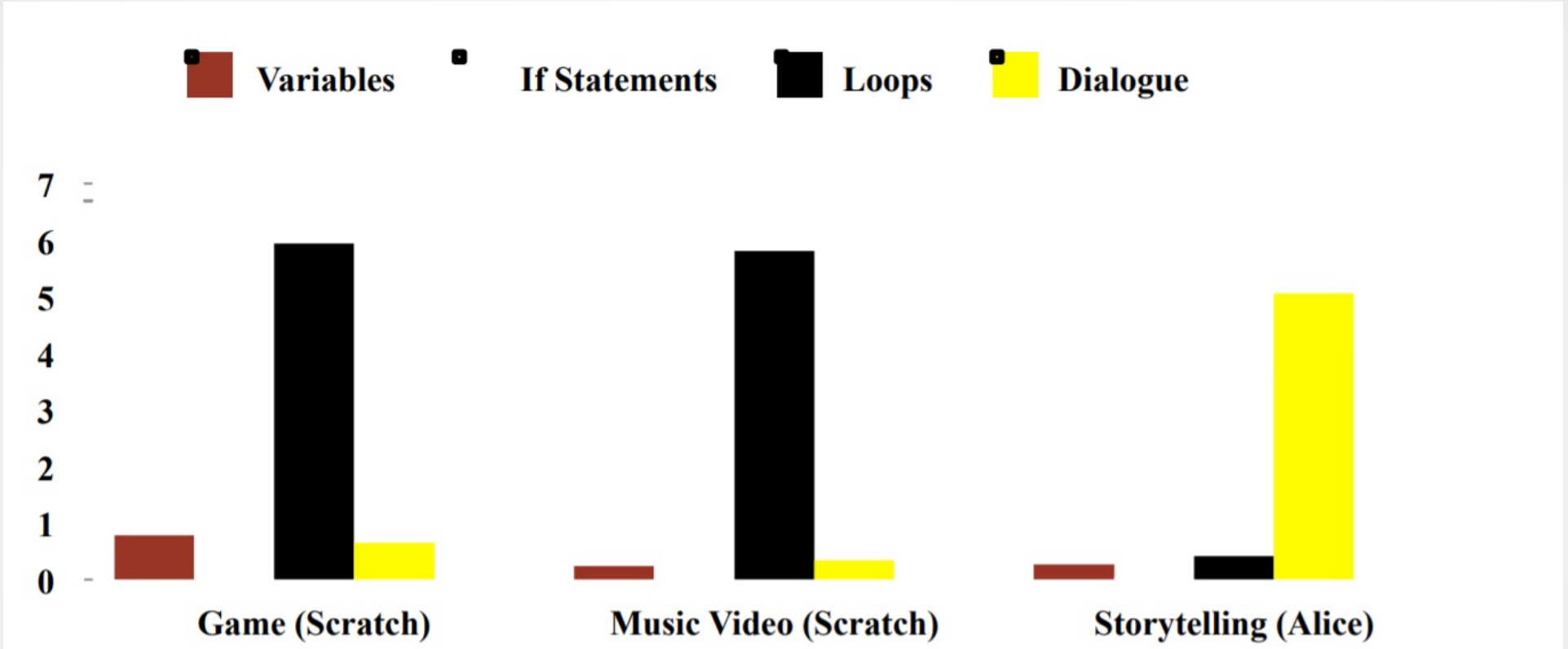


Figure 12. Use of Programming Constructs in Game, Music Video, and Storytelling Projects (Per 100 Lines of Code)

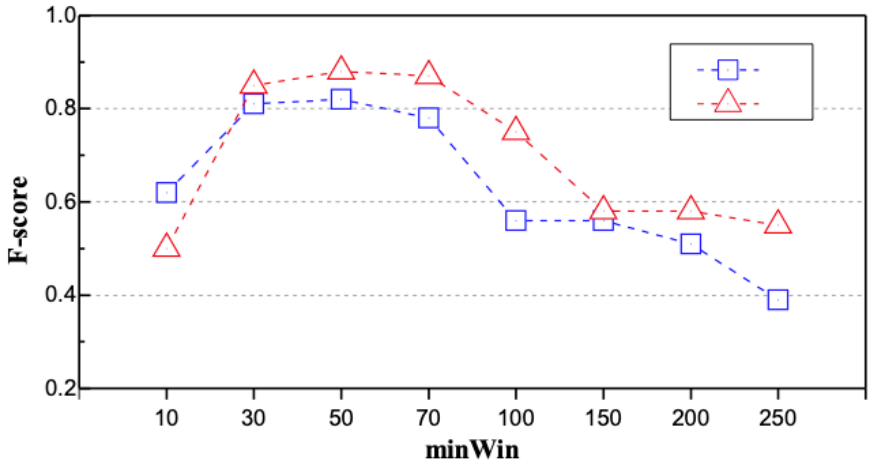


Fig. 14. F-scores obtained with different minWin

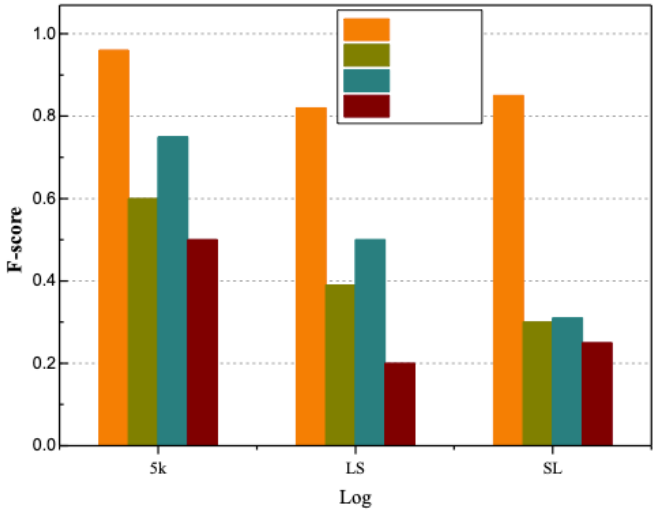
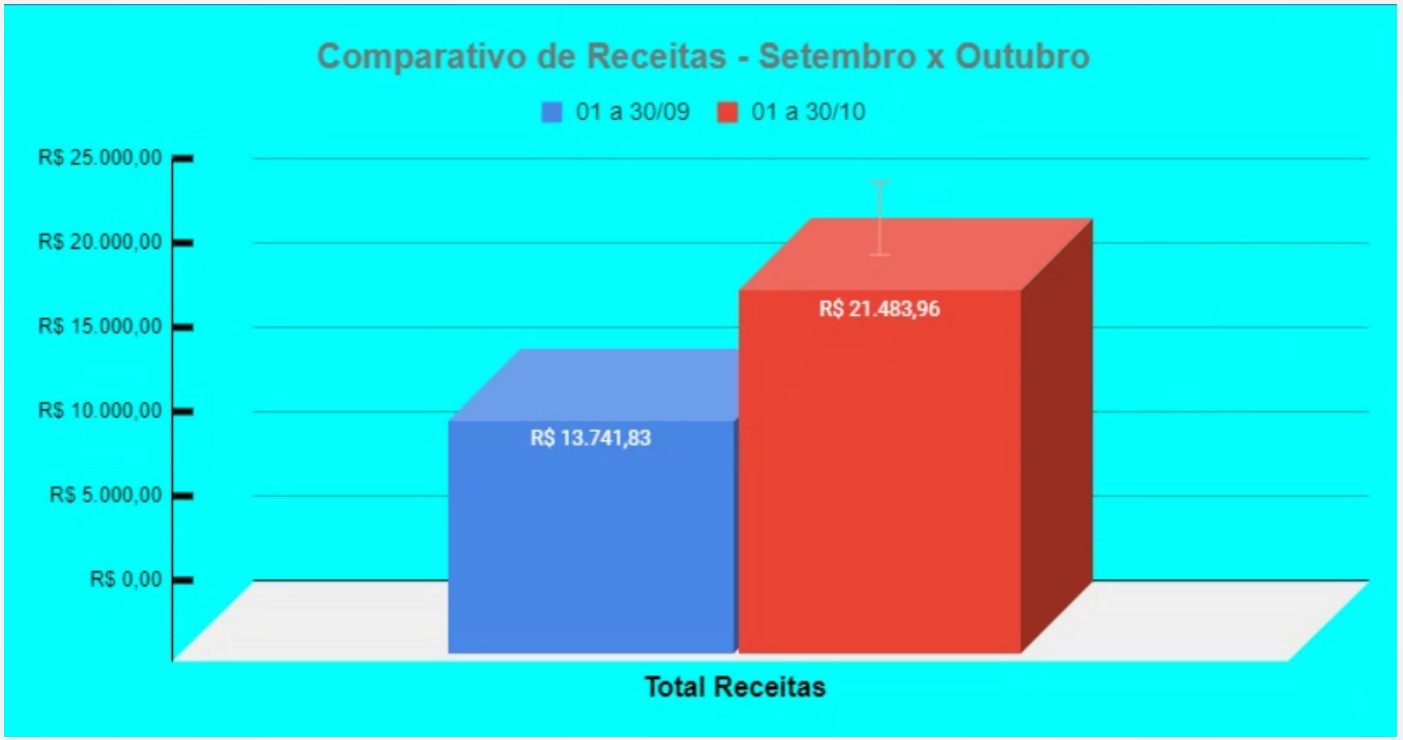


Fig. 15. F-scores obtained under different adaptive window strategies


Comparativo de Receitas - Setembro x Outubro
01 a 30/09
01 a 30/10
R$ 25.000,00
R$ 20.000,00
R$ 15.000,00
R$ 10.000,00
R$ 5.000,00
R$ 0,00
R$ 13.741,83
R$ 21.483,96
Total Receitas

References

Most of these visualizations were obtained from

https://badvisualisations.tumblr.com/

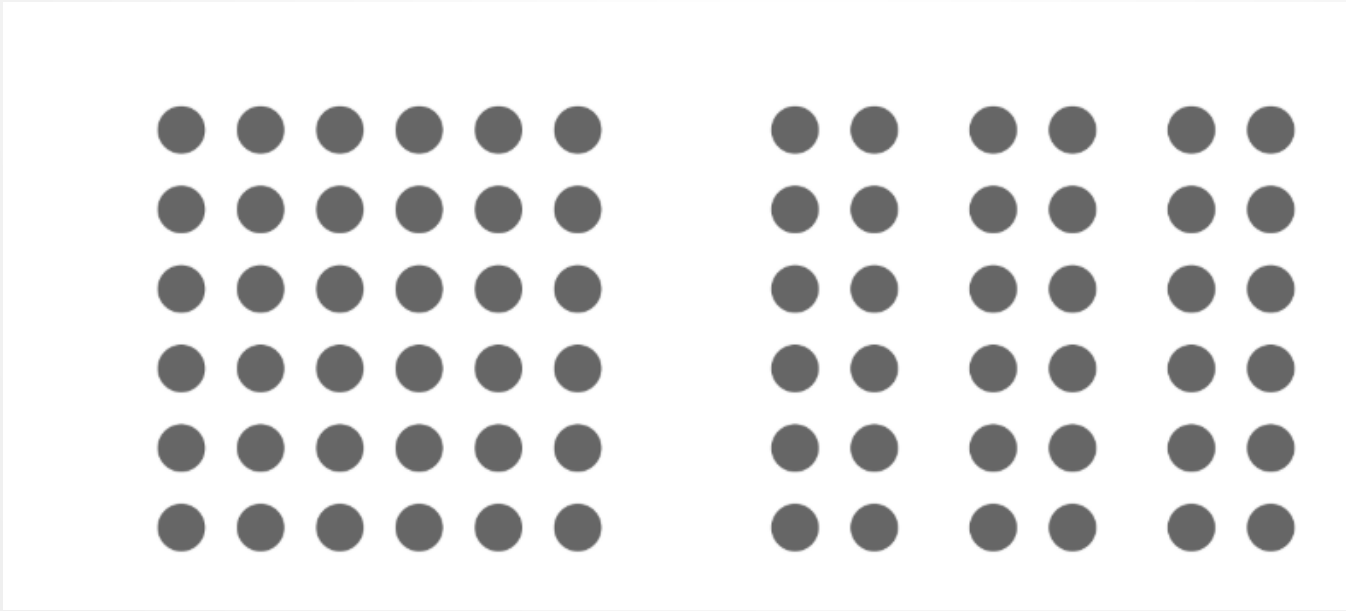
GESTALT PRINCIPLES

Gestalt principles

- Proximity
- Similarity
- Enclosure
- Closure
- Continuity
- Connection

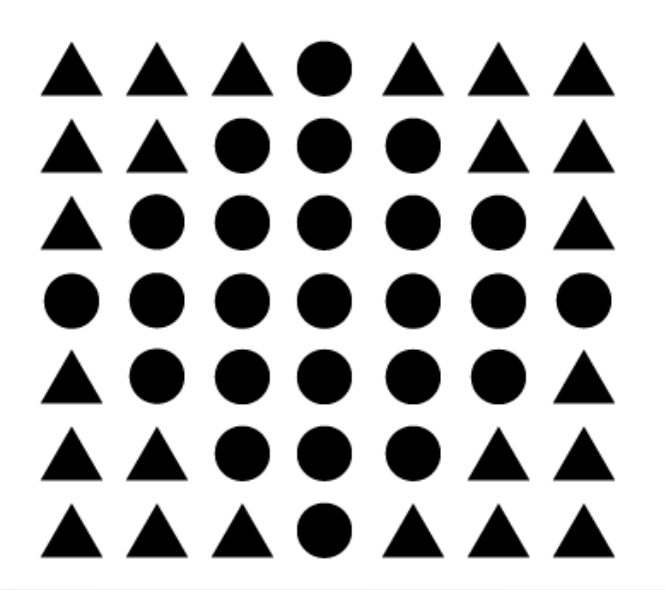
Proximity

- Things that are closer to one another are perceived as belonging to a same group



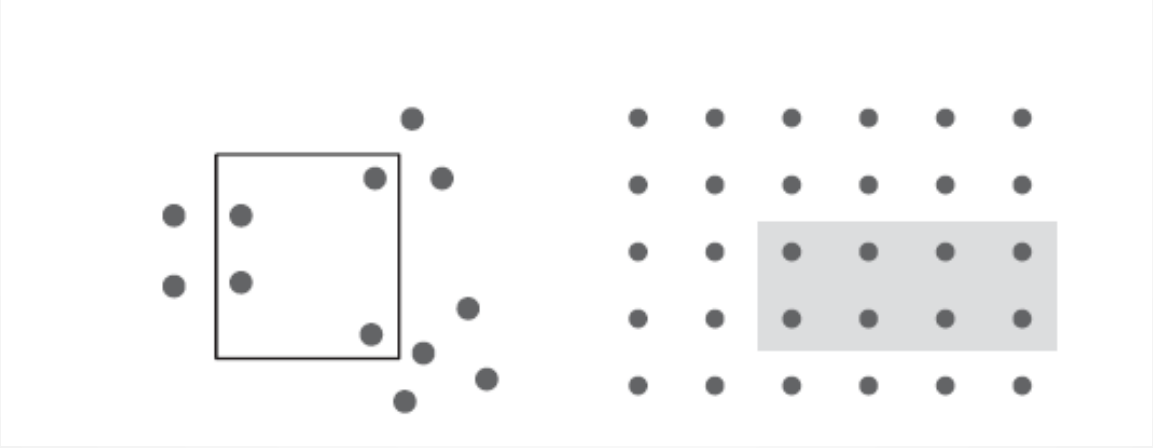
Similarity

- Objects that share shapes, sizes, colors or orientation are perceived as belonging to the same group



Enclosure

- We observe objects that are enclosed together as belonging to the same group



Closure

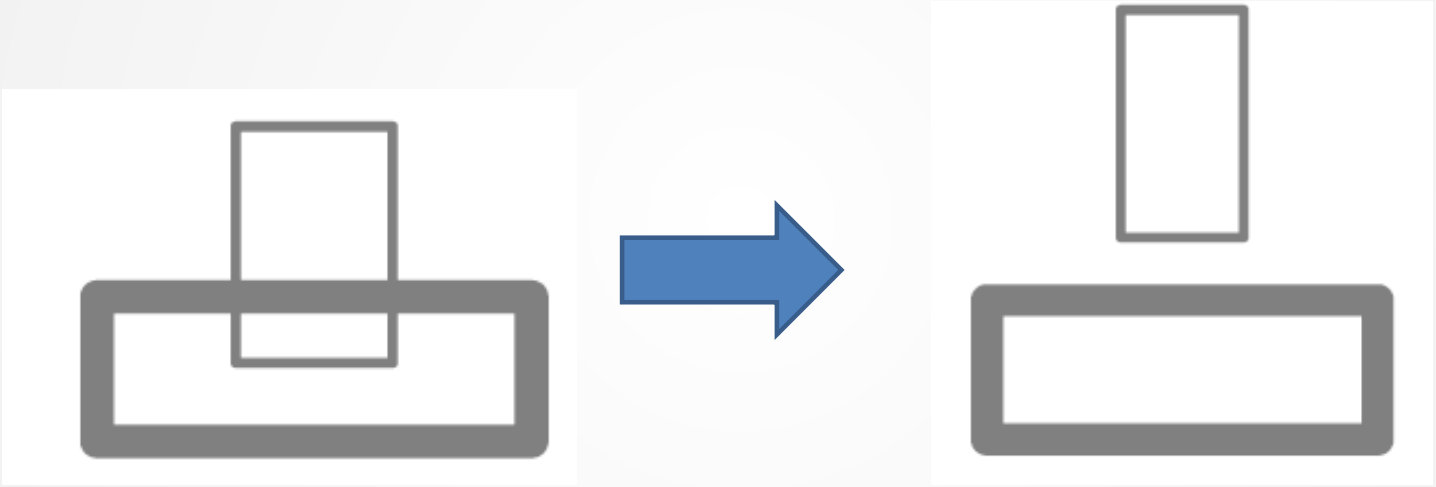
- We perceive objects as a whole even though some parts are missing.

Continuity

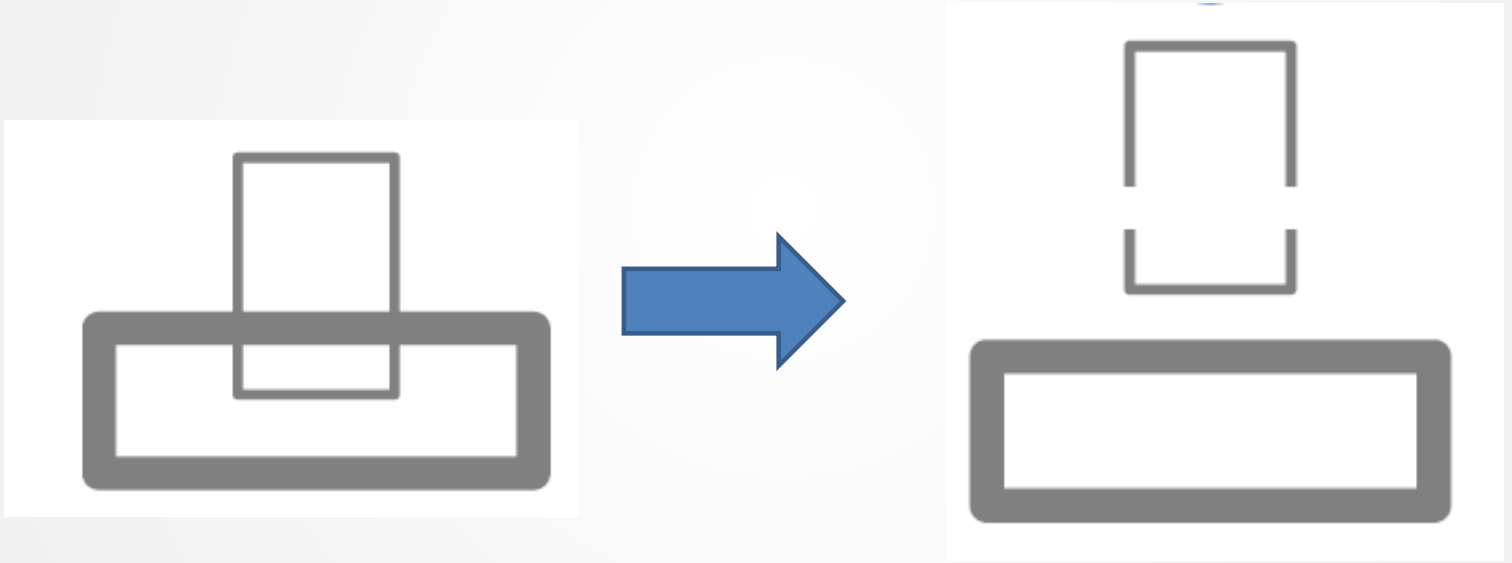
- Our eyes seek the most “natural” and “smooth” path between objects, even though they may not exist

How would you separate these items?

Continuity - 1



Continuity - 2



Connection

- We tend to perceive connected objects as a group

- Connection is often stronger than color, shape or size

Analyze the following visualization (5 minutes).

Next, you will be asked a (tough) question about it and the gestalt principles.

Market size over time

Which Gestalt principles have been used?

Which gestalt principles have been used?

- Proximity:
  - Indicates that the y axis, title and labels must be read together
  - Clarifies that the data labels and markers are related

- Similarity:
  - The similarity of colors (orange and blue) with the text is used to connect things

Which gestalt principles have been used?

- Enclosure:
  - The gray region is used to differentiate the forecasts from the historical values

Which gestalt principles have been used?

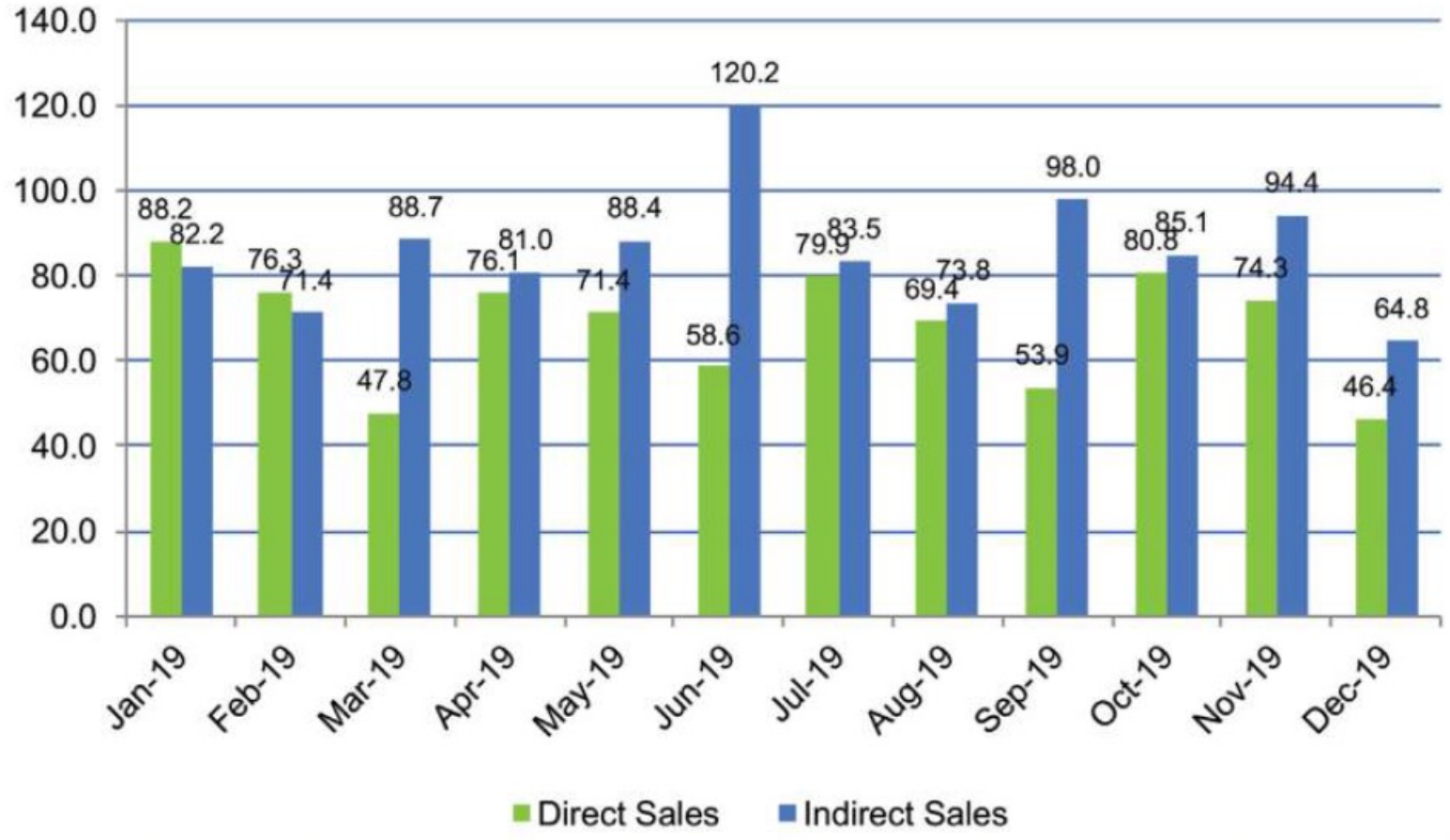
- Continuity:
  - The dashed line is used to connect the forecasts in the right section of the plot

- Connection:
  - In the line plot, all points are connected and make the trend easily visible

Which visual components would you change in the following visualization? (5 min)

Time to Close Deal

Goal = 90 days



1. Removing the external blue lines

- The lines between the title and the plot, as well as the most external line are unnecessary

- The enclosure principle allows us to visualize the plot without them


Time to Close Deal
Goal = 90 days
140.0
120.0
100.0
80.0
60.0
40.0
20.0
0.0
88.2
82.2
76.3
71.4
47.8
88.7
76.1
81.0
71.4
88.4
58.6
120.2
79.9
83.5
69.4
73.8
53.9
98.0
80.8
85.1
74.3
94.4
46.4
64.8
Jan-19
Feb-19
Mar-19
Apr-19
May-19
Jun-19
Jul-19
Aug-19
Sep-19
Oct-19
Nov-19
Dec-19
Direct Sales
Indirect Sales

2. Remove the grid lines

- Removing the grid lines, our attention is drawn to the data


Time to Close Deal
Goal = 90 days
140.0
120.0
100.0
80.0
60.0
40.0
20.0
0.0
88.2
82.2
76.3
71.4
47.8
88.7
76.1
81.0
71.4
88.4
58.6
120.2
79.9
83.5
69.4
73.8
53.9
98.0
80.8
85.1
74.3
94.4
46.4
64.8
Jan-19
Feb-19
Mar-19
Apr-19
May-19
Jun-19
Jul-19
Aug-19
Sep-19
Oct-19
Nov-19
Dec-19
Direct Sales
Indirect Sales

3. Remove the zeroes from the y axis

- The extra zeroes in the decimal places are not required

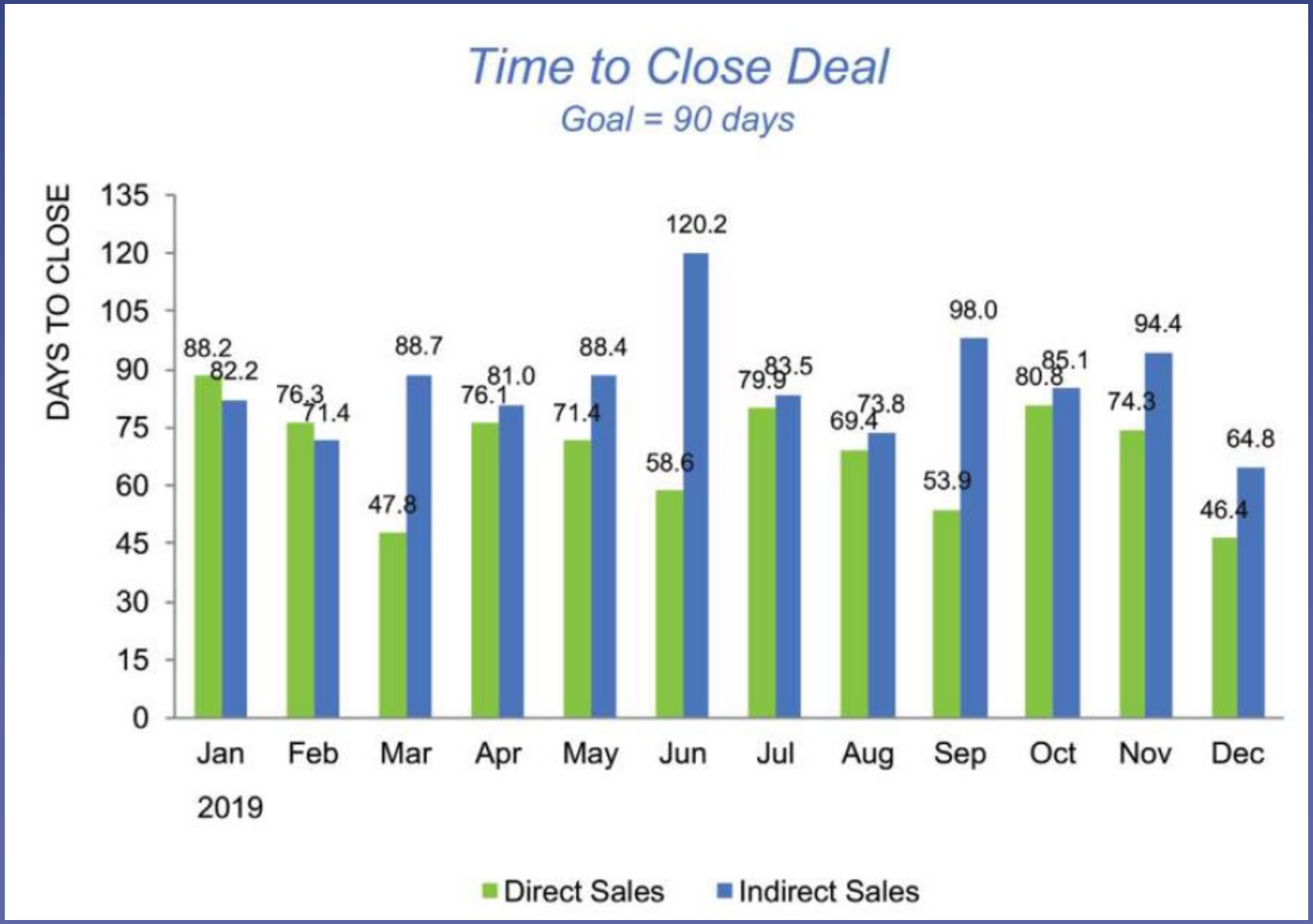
- It is also interesting to change the y axis scale for 15-day intervals


Time to Close Deal
Goal = 90 days
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
88.2
82.2
76.3
71.4
47.8
88.7
76.1
81.0
71.4
88.4
58.6
120.2
79.9
83.5
69.4
73.8
53.9
98.0
80.8
85.1
74.3
94.4
46.4
64.8
Jan-19
Feb-19
Mar-19
Apr-19
May-19
Jun-19
Jul-19
Aug-19
Sep-19
Oct-19
Nov-19
Dec-19
Direct Sales
Indirect Sales

4. Eliminate diagonal texts in the x axis

- Diagonal and vertical texts are polemic

- Whenever possible, prefer horizontal texts


Time to Close Deal
Goal = 90 days
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
88.2
82.2
76.3
71.4
47.8
88.7
76.1
81.0
71.4
88.4
58.6
120.2
79.9
83.5
69.4
73.8
53.9
98.0
80.8
85.1
74.3
94.4
46.4
64.8
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019
Direct Sales
Indirect Sales

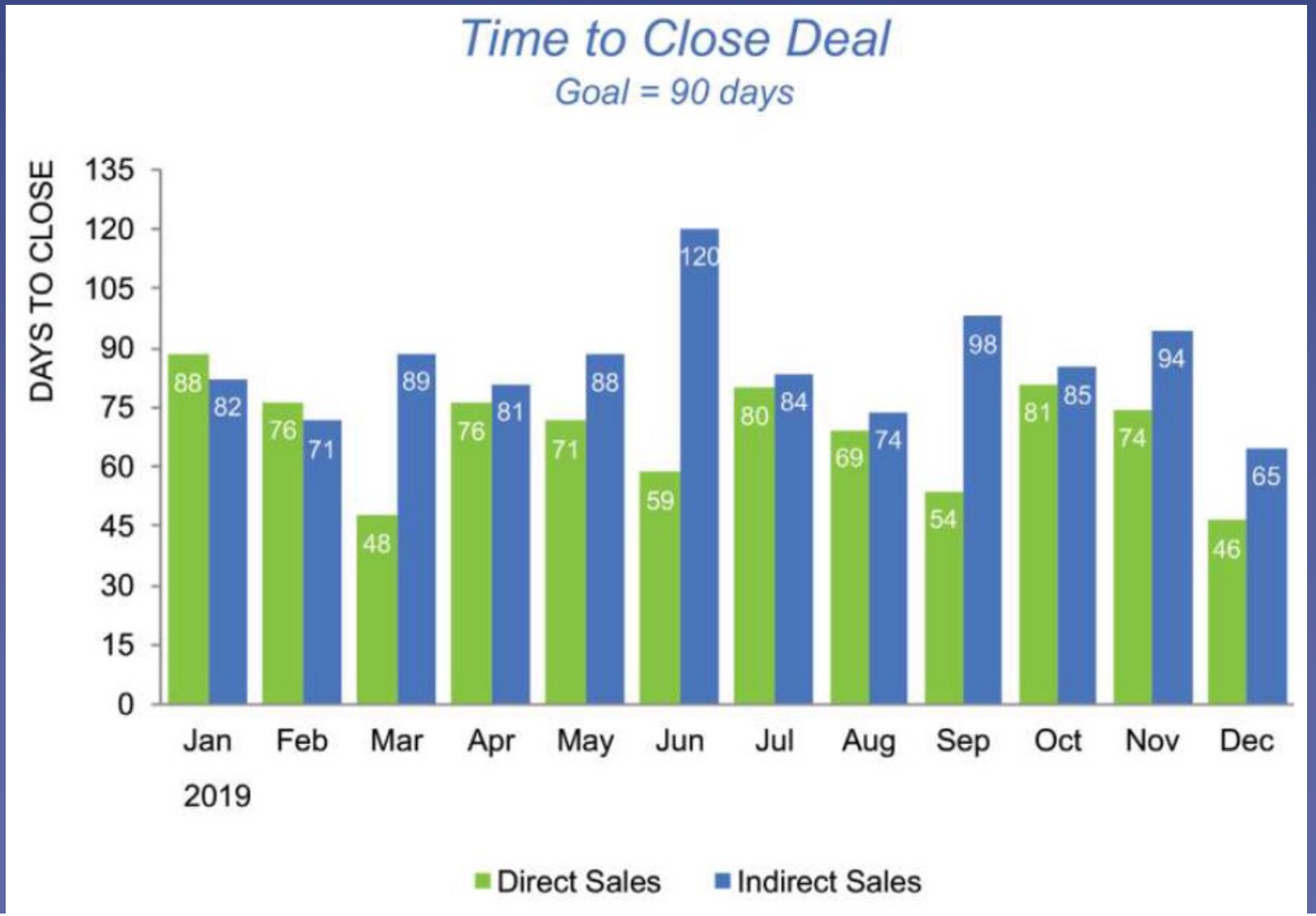
5. Decrease blank spaces

- Avoid having unnecessarily big blank spaces between bars
- Useful due to the connection principle
- A good practice, however, is to keep blank spaces between bars from different categories


Time to Close Deal
Goal = 90 days
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
88.2
82.2
76.3
71.4
47.8
88.7
76.1
81.0
71.4
88.4
58.6
120.2
79.9
83.5
69.4
73.8
53.9
98.0
80.8
85.1
74.3
94.4
46.4
64.8
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019
Direct Sales
Indirect Sales

6. “Drag” the labels to the bars

- Whenever possible, round the values


Time to Close Deal
Goal = 90 days
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
88
82
76
71
48
89
76
81
71
88
59
120
80
84
69
74
54
98
81
85
74
94
46
65
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019
Direct Sales
Indirect Sales

7. Eliminate the data labels

- The y axis is redundant with the numbers provided in the labels

- Important: remove or not to remove?
  - It depends on the context:
    - The exact values are required?
    - Or the trend is more relevant?


Time to Close Deal
Goal = 90 days
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019
Direct Sales
Indirect Sales

8. Make it a line plot


Time to Close Deal
Goal = 90 days
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019
Direct Sales
Indirect Sales

9. Apply the legend to the plot

- Using the proximity principle, we can label our lines inside the plot itself


Time to Close Deal
Goal = 90 days
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019
Indirect
Direct

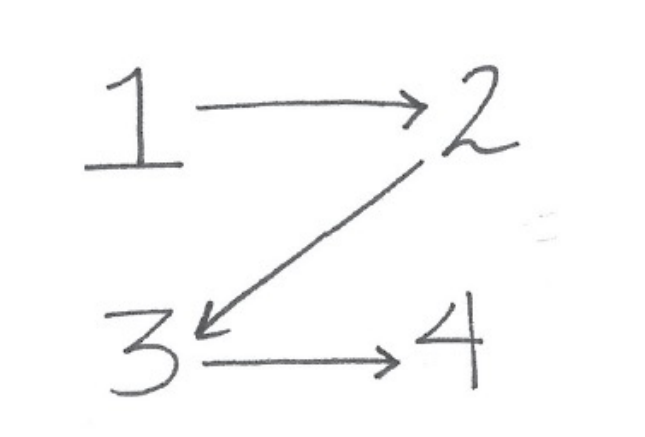
10. Changing the legend color to adhere to the data

- Proximity and similarity


Time to Close Deal
Goal = 90 days
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019
Indirect
Direct

11. Title position

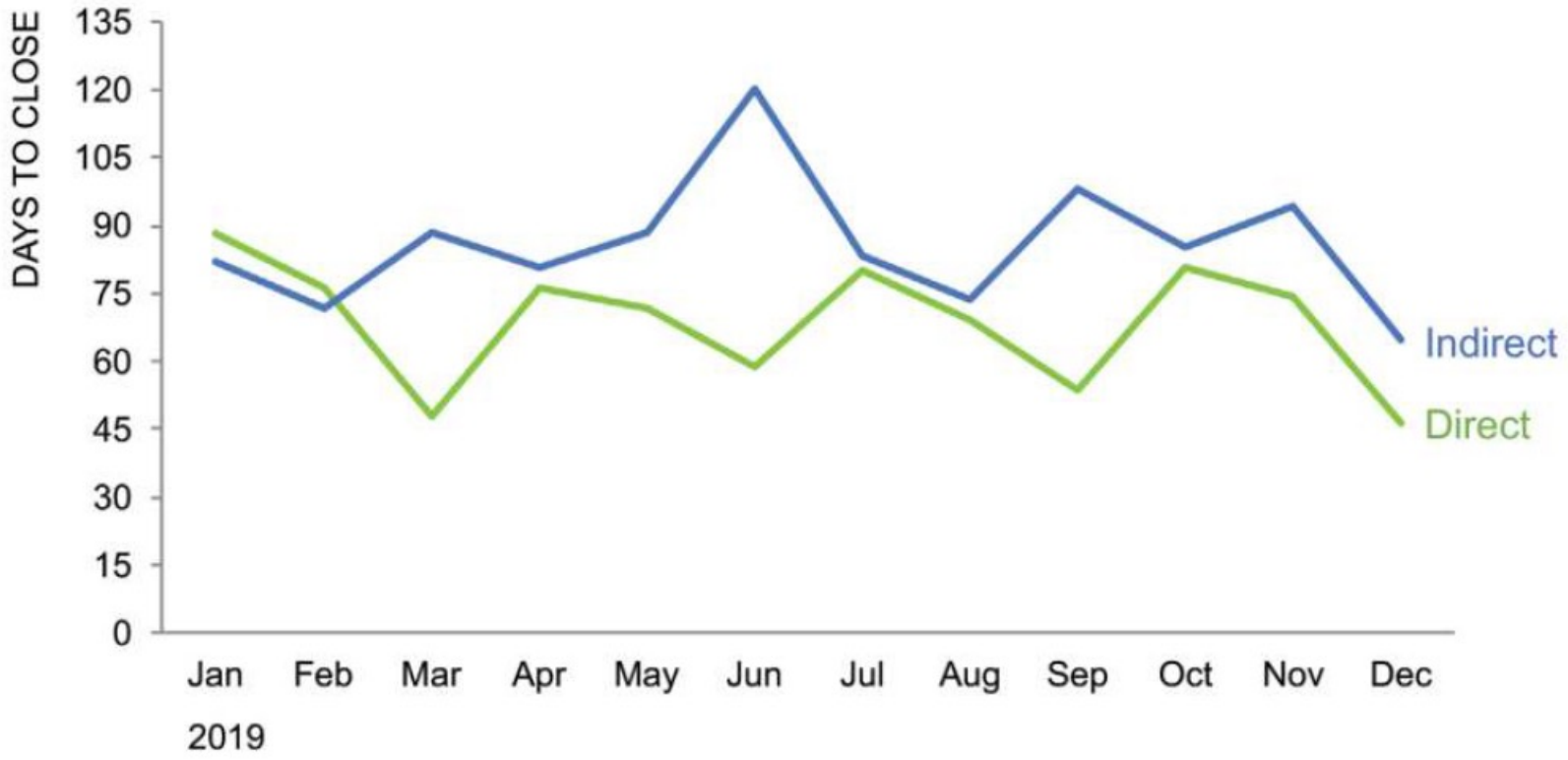
- Do not forget how we read (*zigzagging z's*):



- With this small change, the reader will focus on the title before anything else

Time to Close Deal

Goal = 90 days

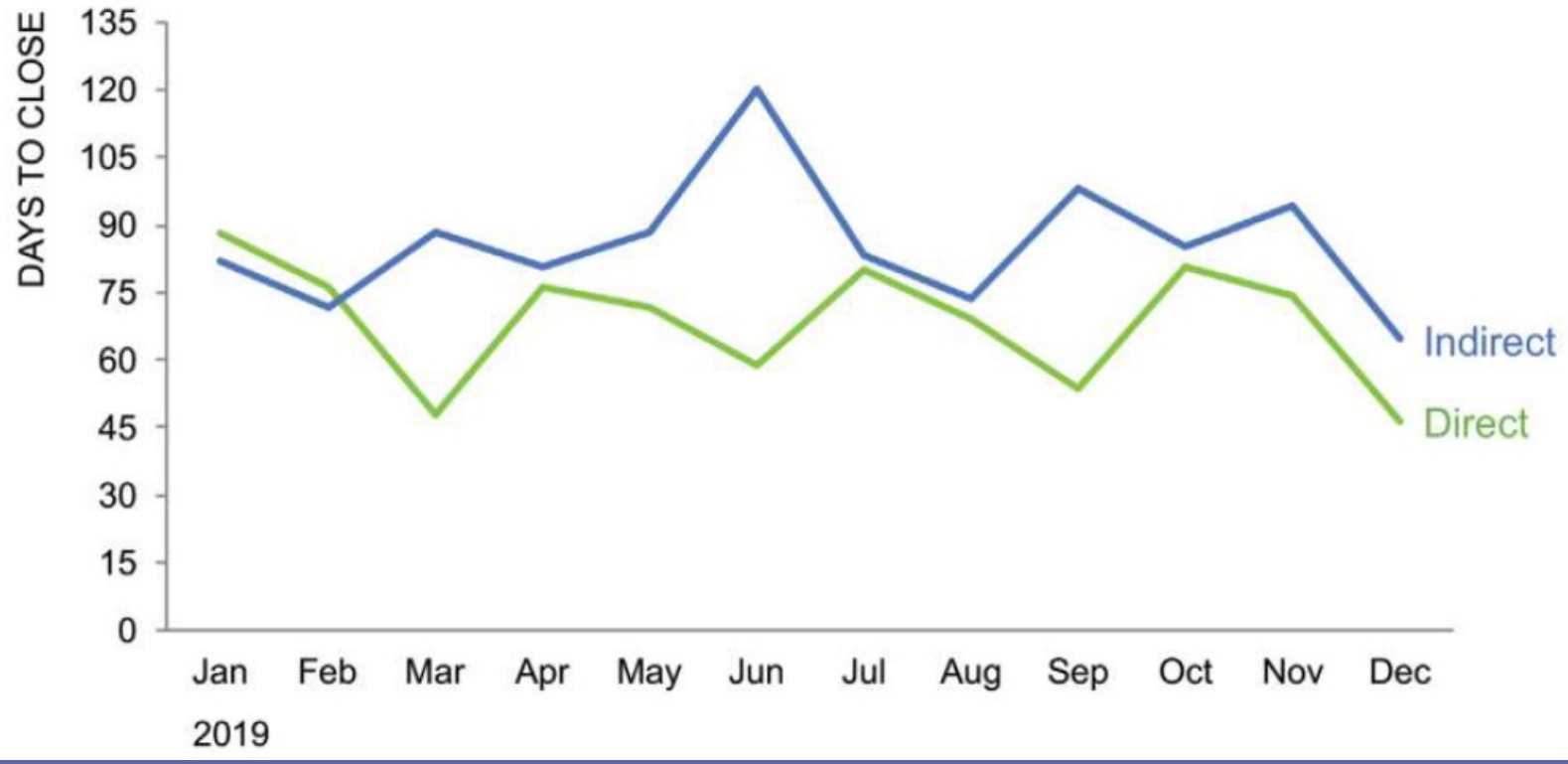


12. Removing the title color

- Is the color from the title anyhow related to the “indirect” component?

Time to close deal

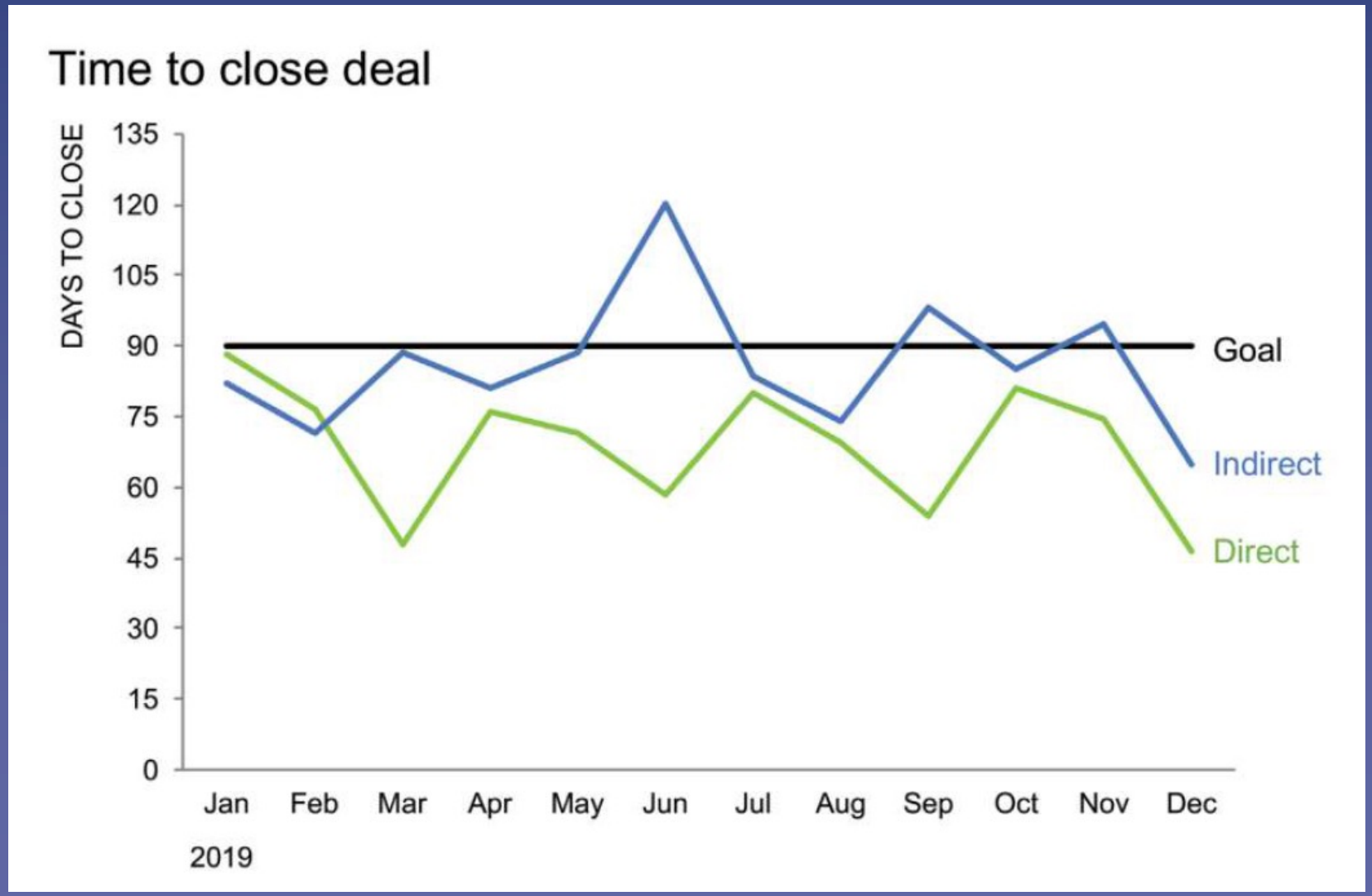
Goal = 90 days



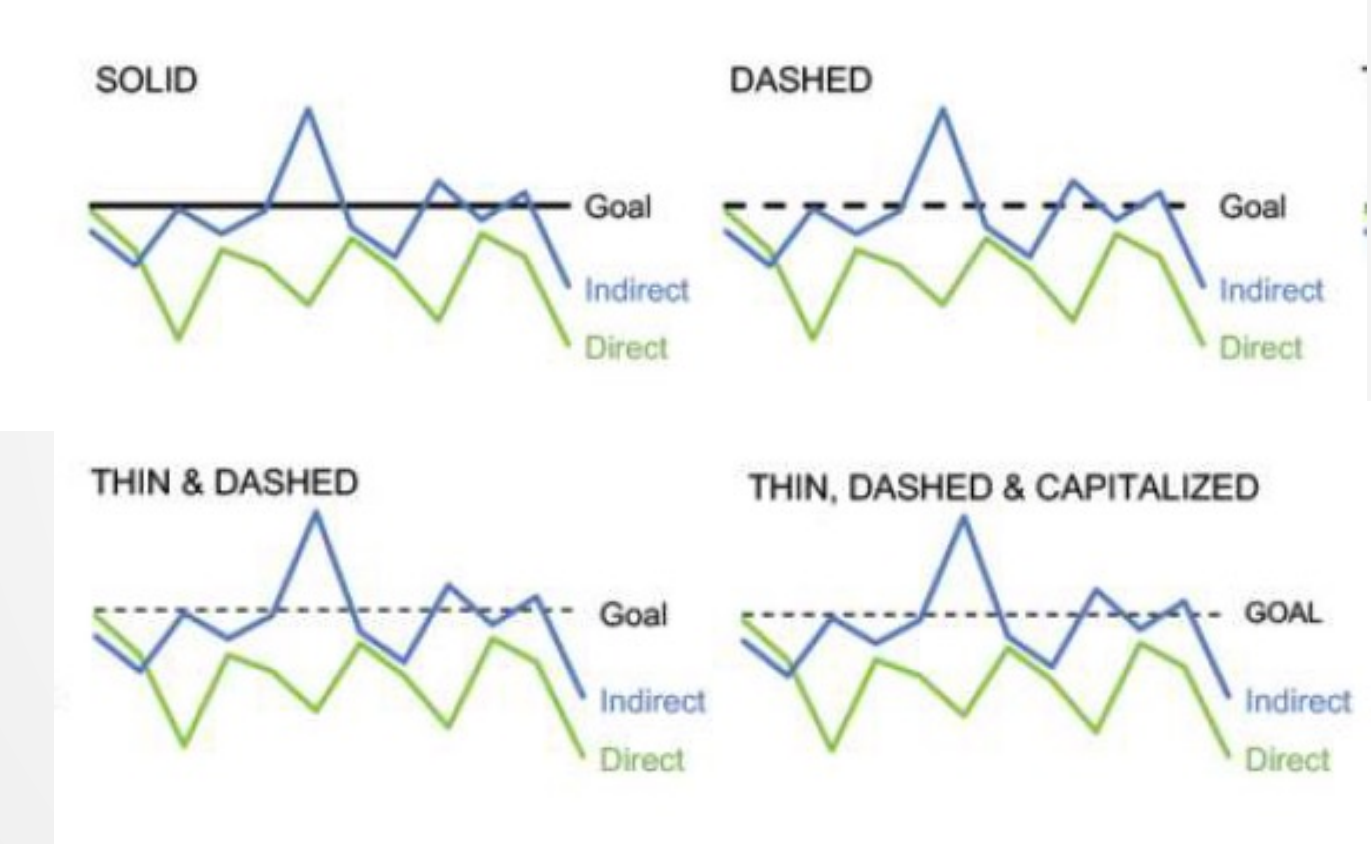
13. Adding the goal to the plot

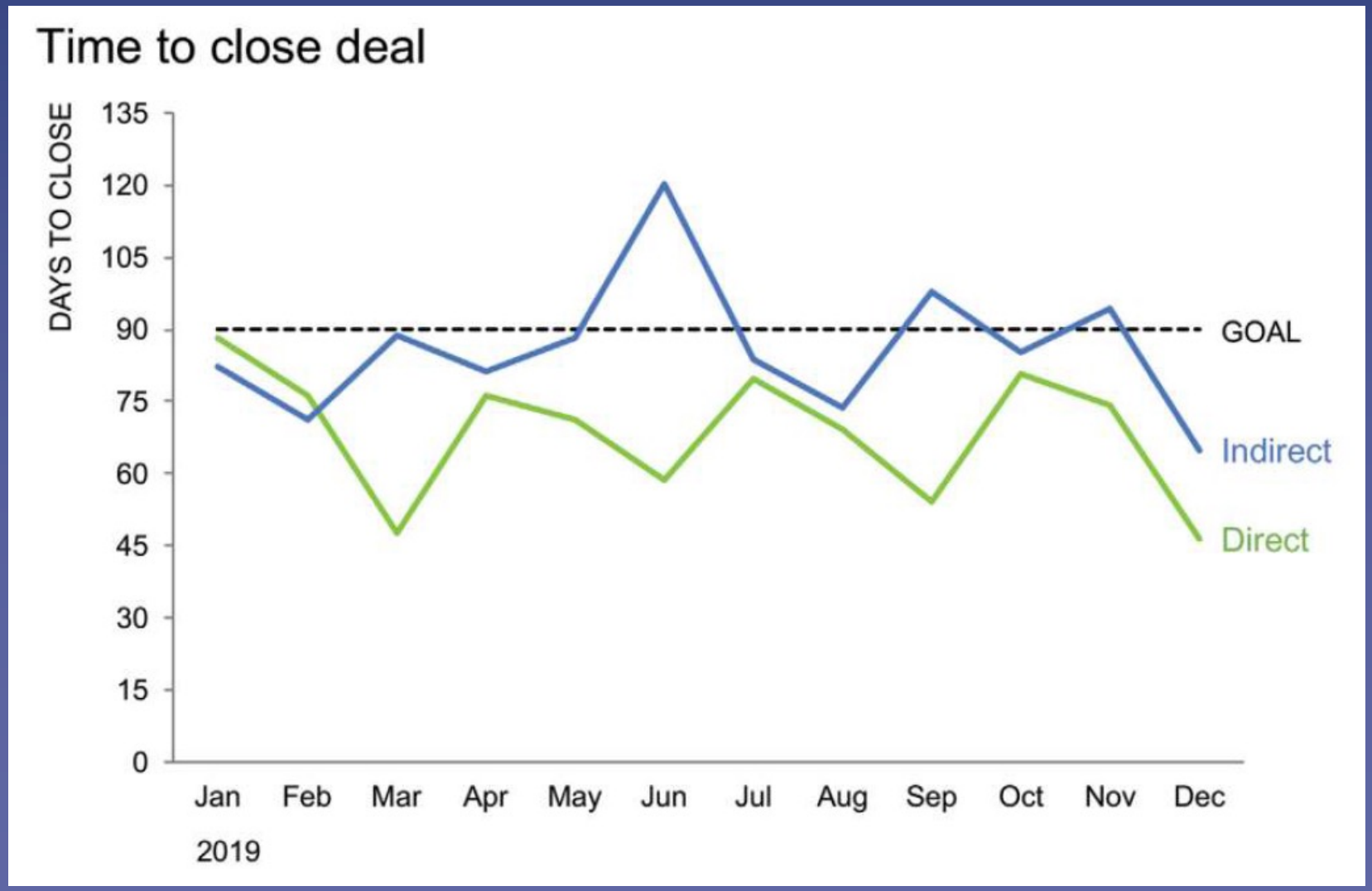
- The goal to make a deal is 90 days

- This infromation can be added to the plot to make the analysis of the curves more visual


Time to close deal
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019
Goal
Indirect
Direct

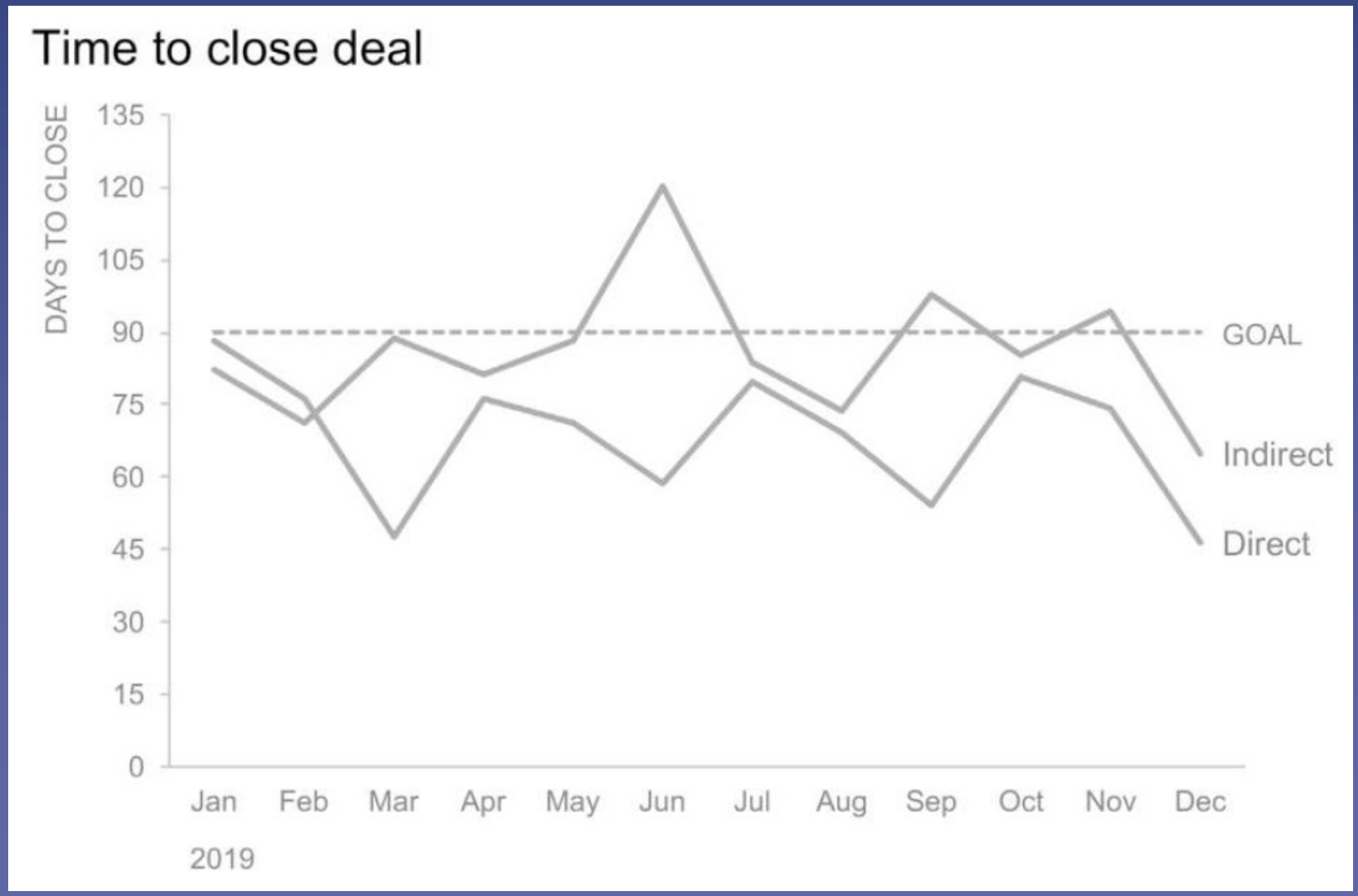
14. Testing different approaches




Time to close deal
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
GOAL
Indirect
Direct
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019

15. Removing colors

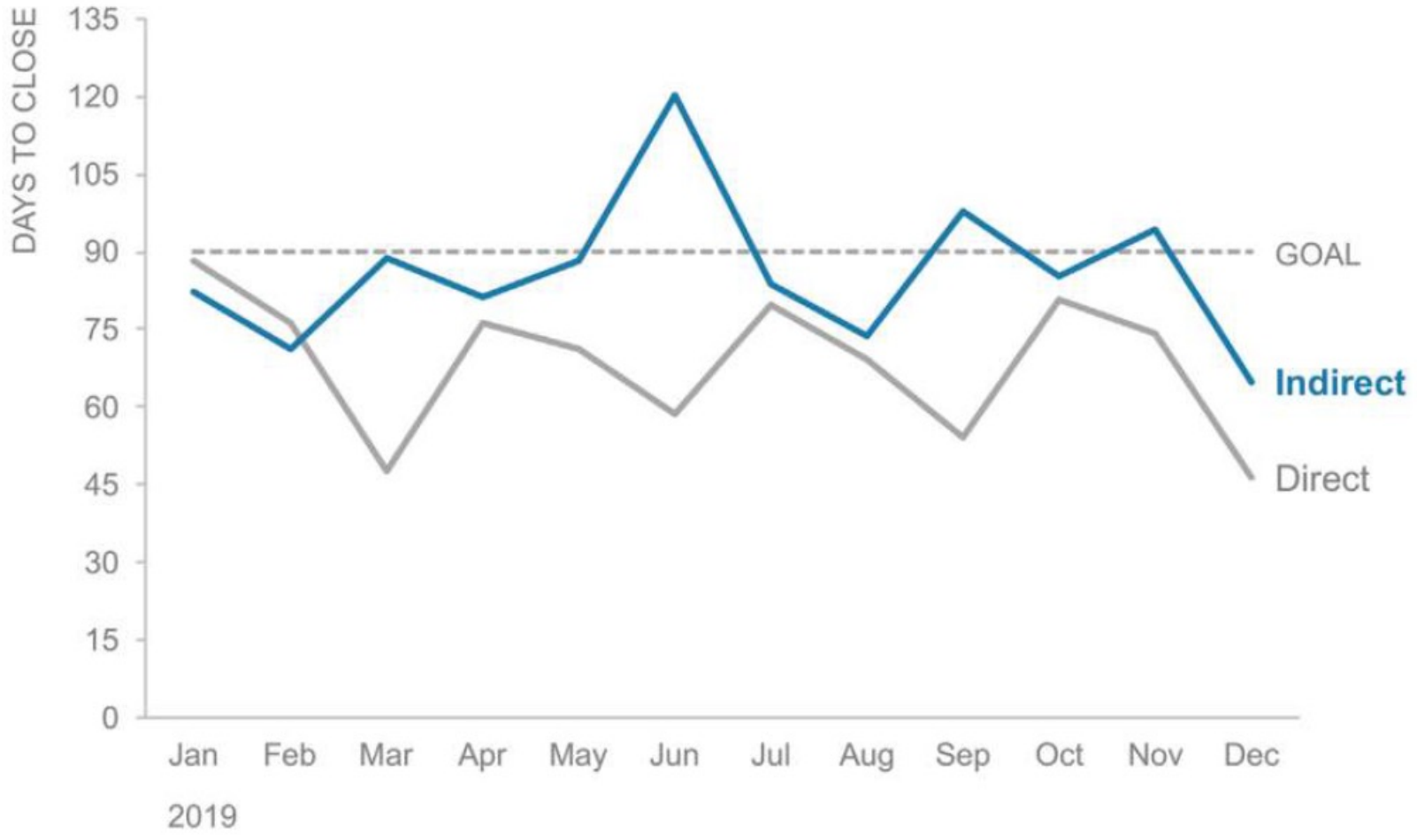
- We have enough separation between the lines


Time to close deal
DAYS TO CLOSE
135
120
105
90
75
60
45
30
15
0
GOAL
Indirect
Direct
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2019

16. Drawing attention

- Depending on the audience and goal of the visualization, we may draw the attention to one of the lines

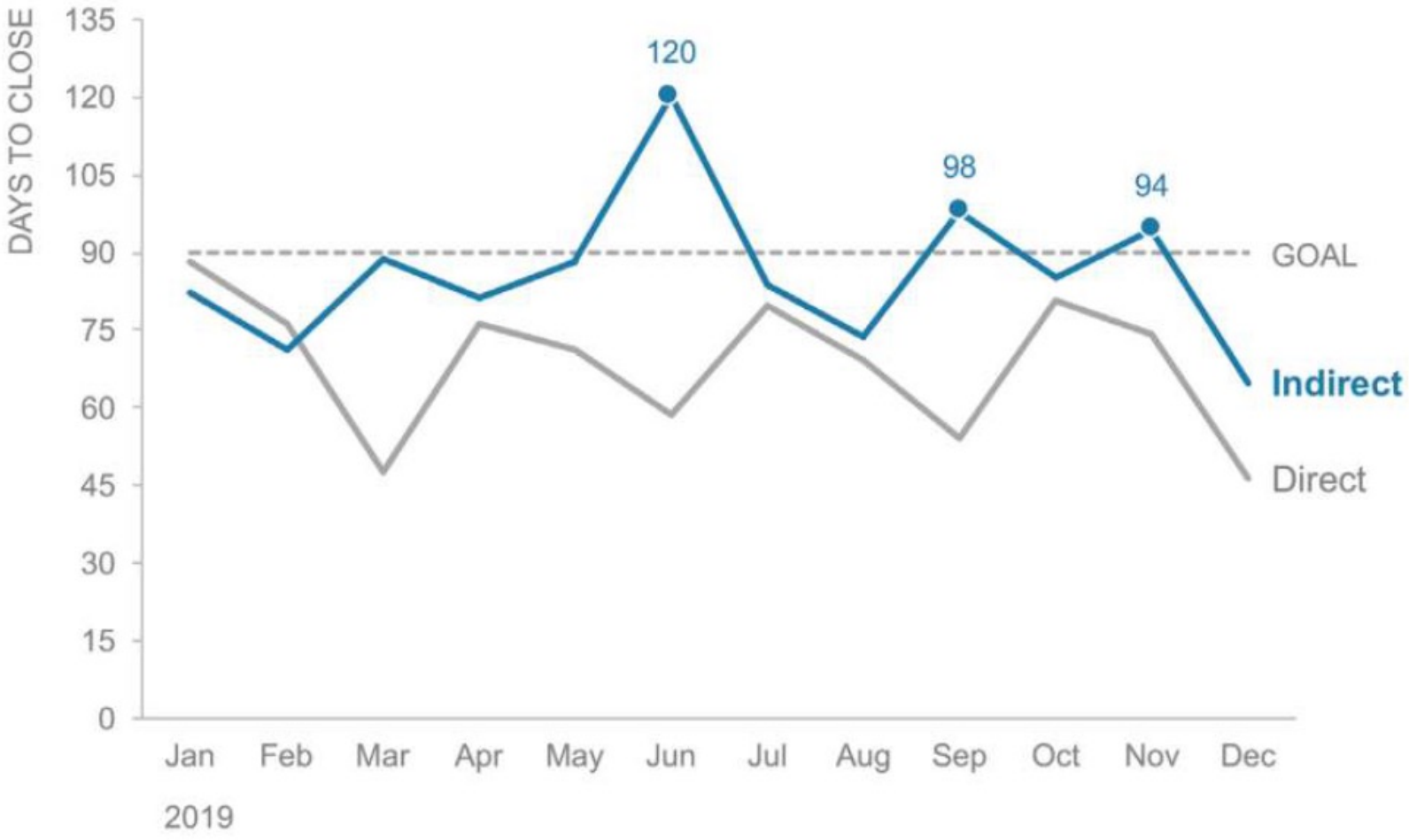
Time to close deal: **indirect varies over time**



17. Focusing on other aspects

- We can focus on other aspects, depending on what we intend to highlight

Time to close deal: **indirect sales missed goal 3 times**

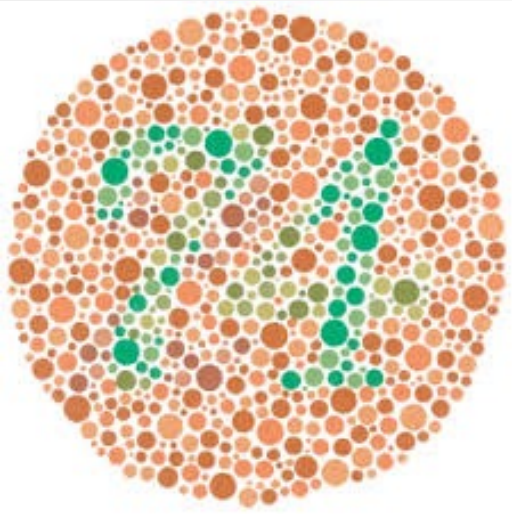


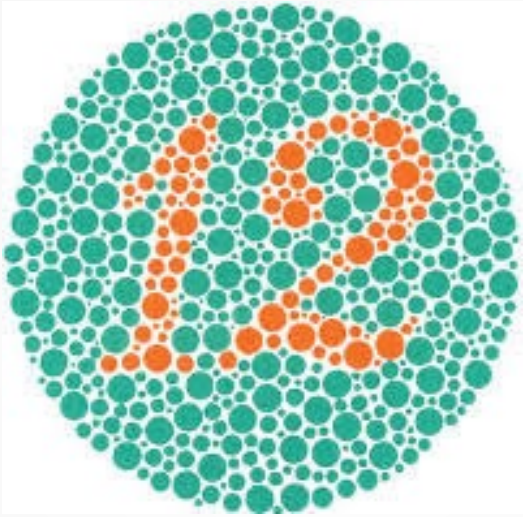
BE MINDFUL OF COLORS

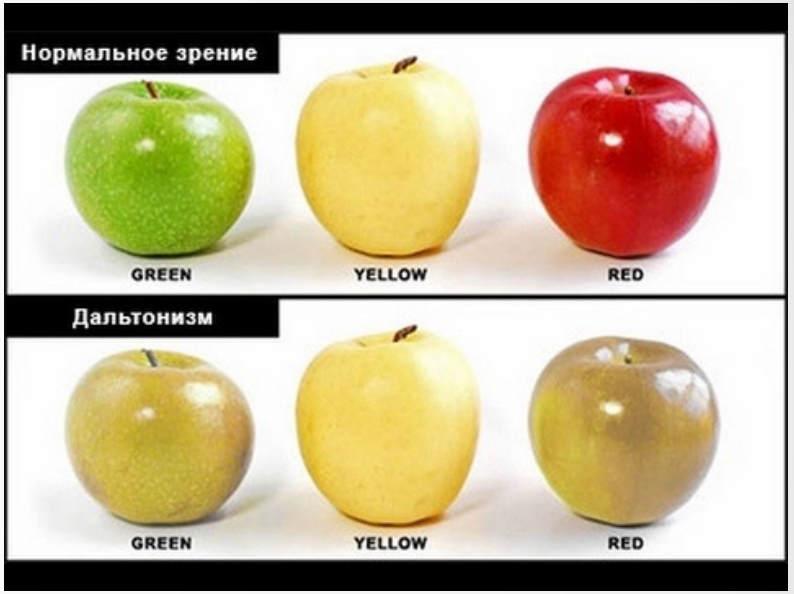
Colors

- One of the most commons mistakes in visualizations regards the poor selection of colors
- Generally, all visualizations should use 2 colors, unless more are indeed needed
- Colors can be used to highlight things
- If colors are needed, avoid intense colors
  - Prefer colors with higher gray values

Color blindness







- Keep in mind: approximately 1 in every 8 men and 1 in every 200 women are colorblind!

Color blindness

- Adobe Color Wheel
  https://color.adobe.com/create/color-accessibility

- Online color blindness test:
  https://enchroma.com/pages/test

- Nice video on how color blindness works:
  https://www.youtube.com/watch?v=iNRQB5309yo

Hints

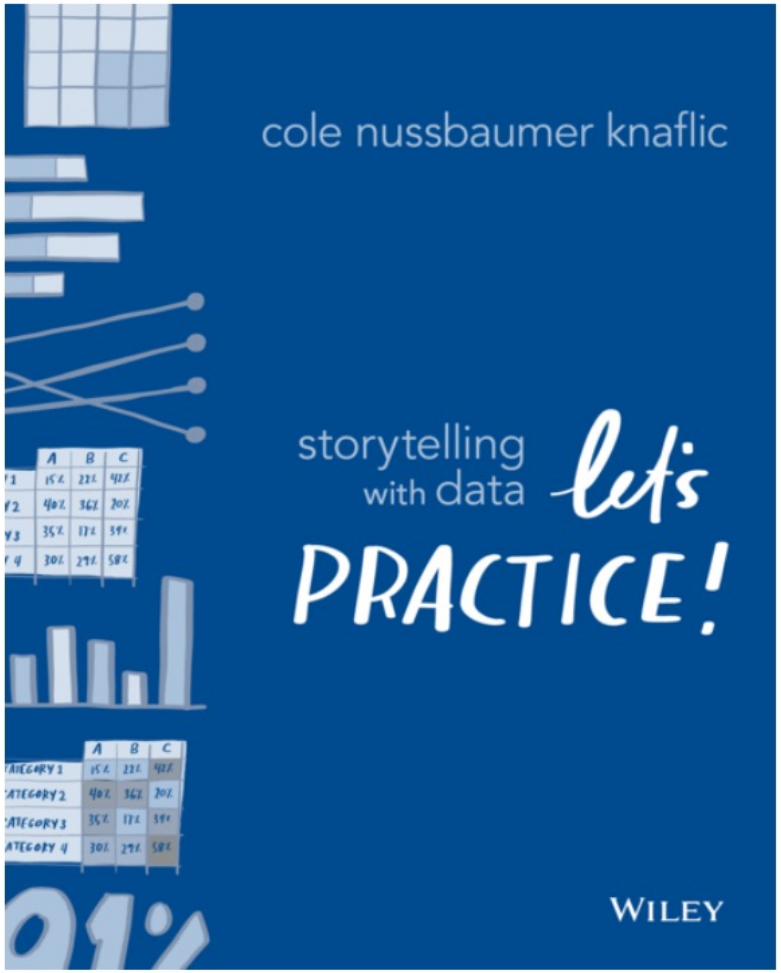
- Avoid using **red colors** and **green** together.
- If you need both together, use another visual component as redundancy
- A suggestion is to use **orange** and **blue**.

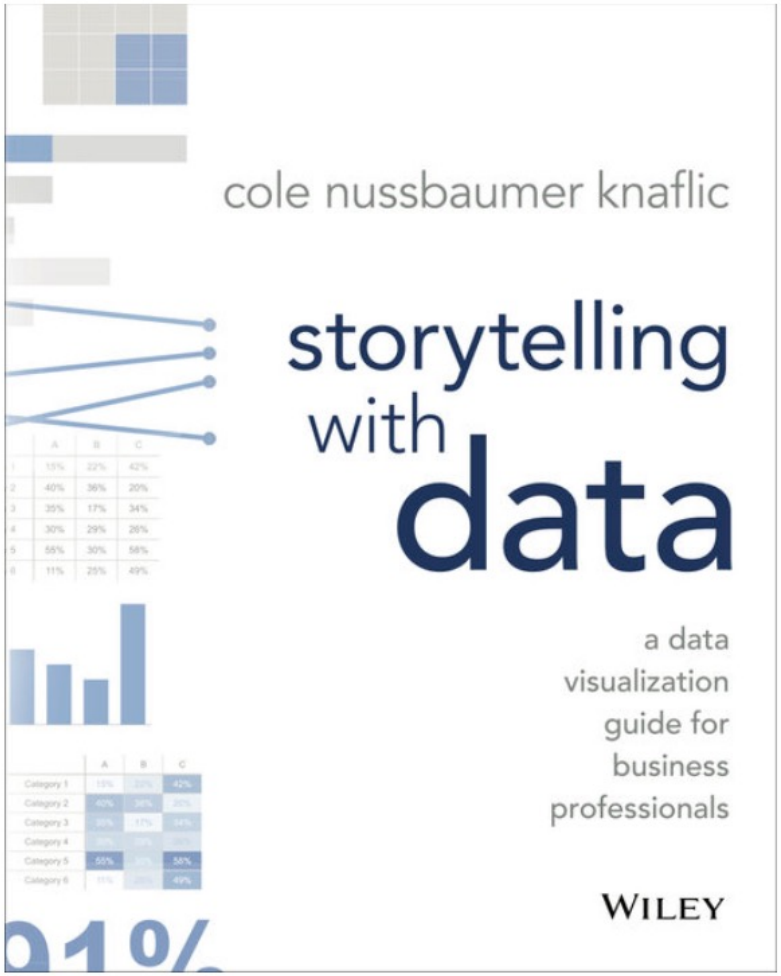
More hints

Avoid the following combinations:

- Green and red
- Green and brown
- Blue and purple
- Green and blue
- Light green and yellow
- Blue and gray
- Green and gray
- Green and black

References


cole nussbaumer knaflic
storytelling with data
let's
PRACTICE!
WILEY


cole nussbaumer knaflic
storytelling with data
a data visualization guide for business professionals
WILEY