


DATA SCIENCE
PPGIA/PUCPR

Prof. Jean Paul Barddal



1

**EXPLORATORY DATA ANALYSIS VERSUS
EXPLANATORY DATA ANALYSIS**

4

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 explanatory analysis
 - Highlights the insights obtained
 - Is often coupled with a story or demand

5

Steps

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

6

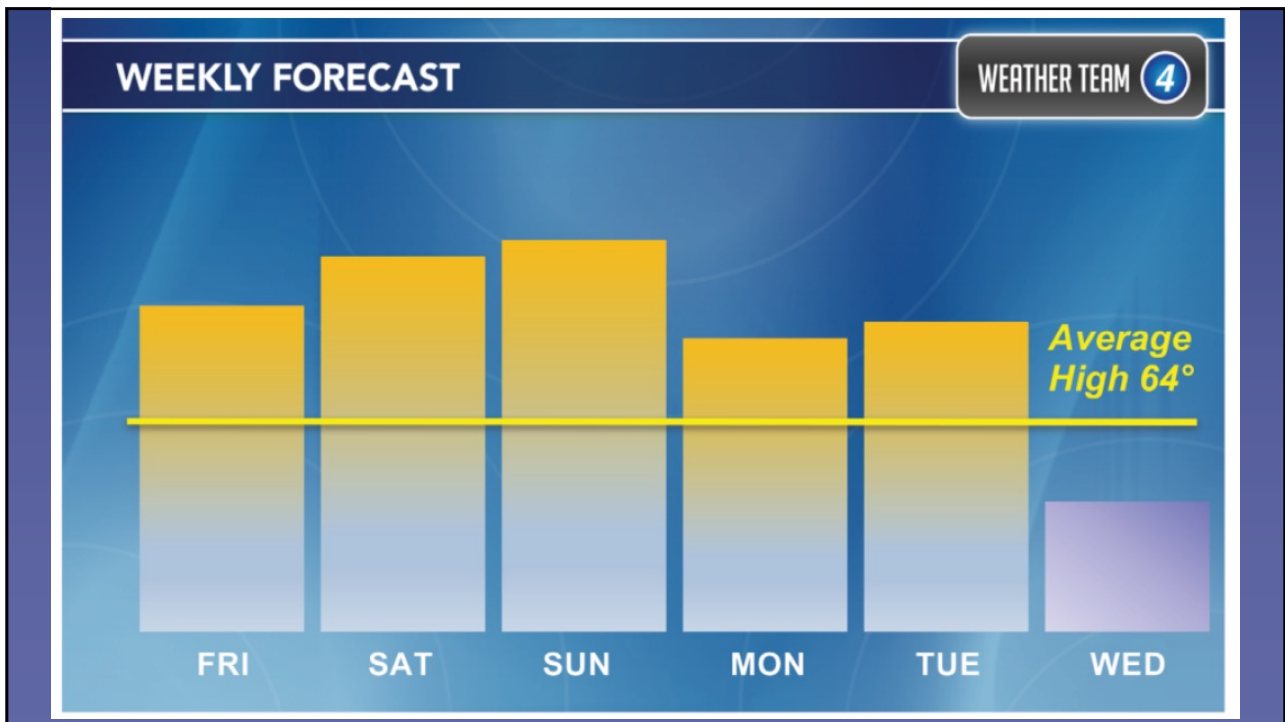
I have a dataset and I need to present it to someone else.

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

7

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

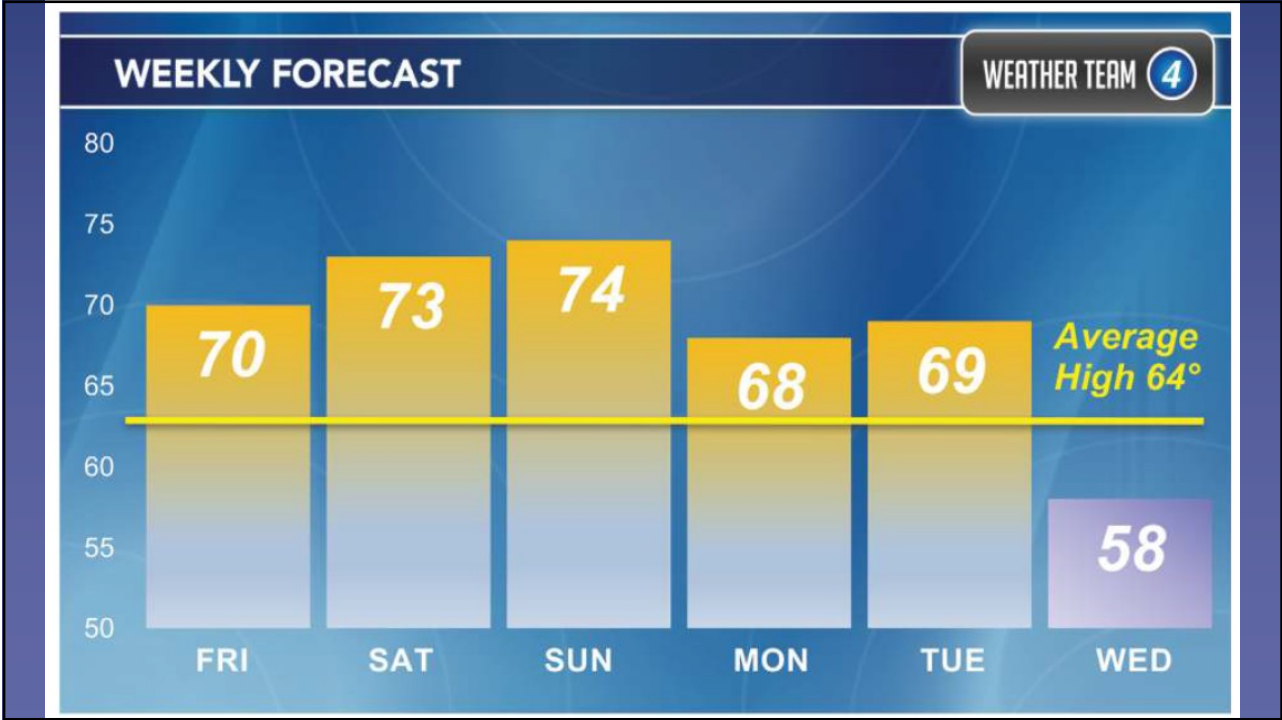
8



9

What temperature would you estimate for Sunday?

10



11



12

EFFECTIVE DATA VIZUALIZATION

13

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 absurdum*” approach in the sense that we will check what should **NOT** be done

14

Checklist

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

15

POOR DATA VISUALIZATION

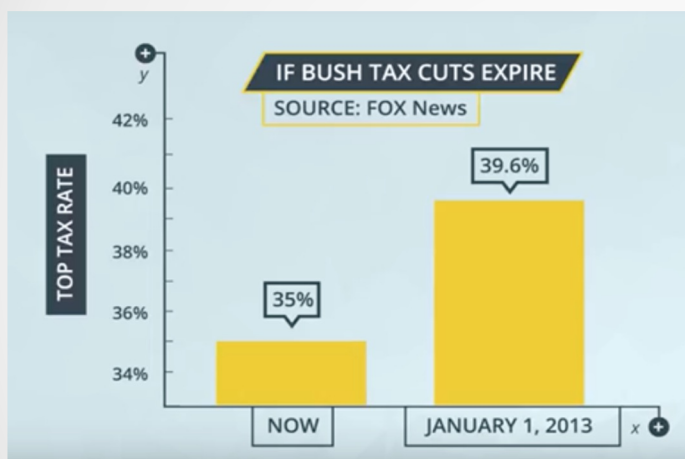
16

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

17

Example

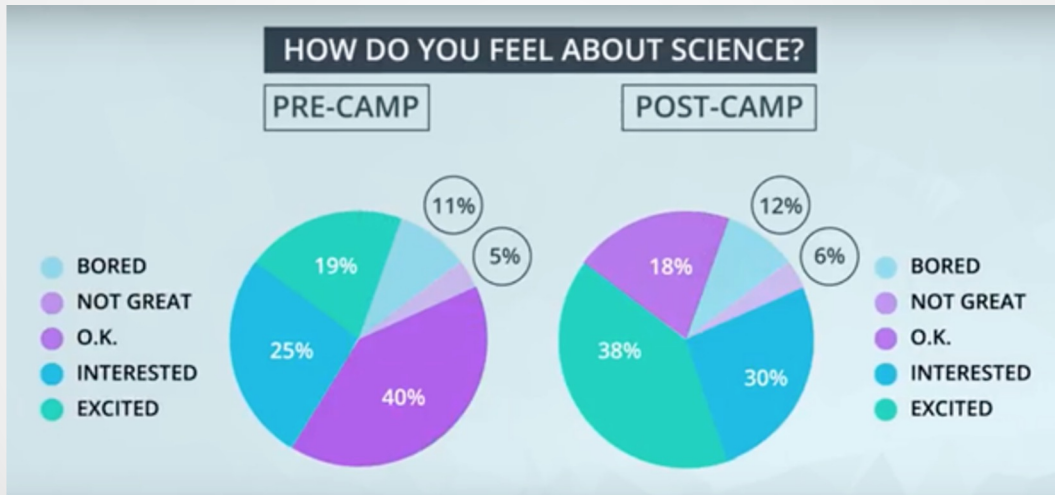


- The difference between the bars is somewhat small, but the scale makes us believe the difference is huge

18

Example

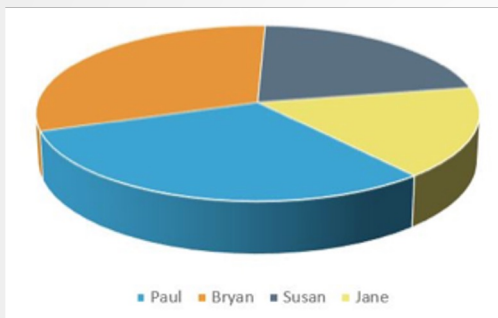
Is it possible to say that the interest in science increased with the camp?



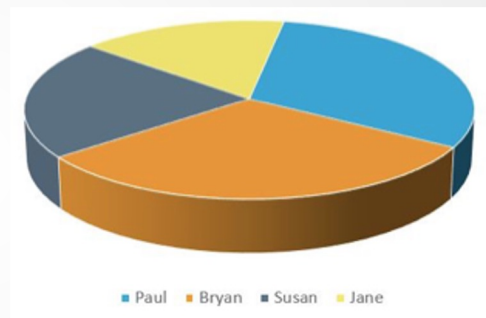
19

Example

Which of the regions below is larger?



Bryan or Paul?

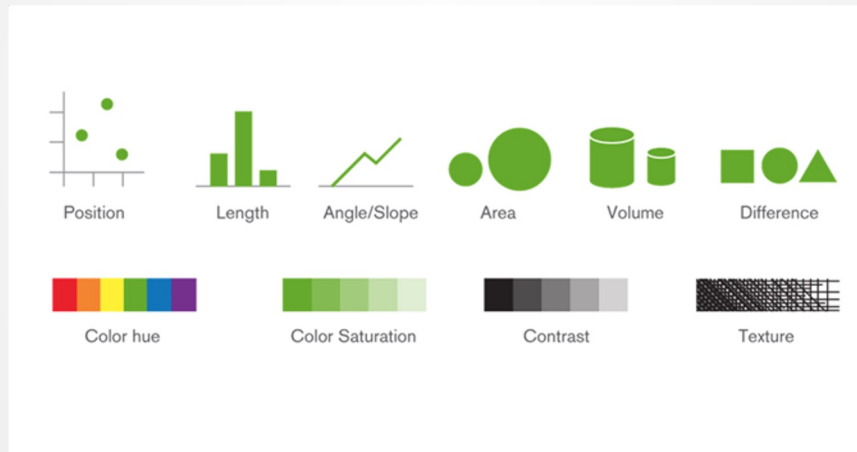


What about here?

20

Visual components

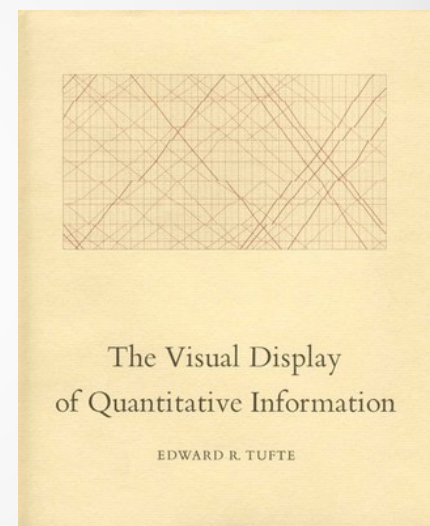
Data visualizations are tailored using the following components:



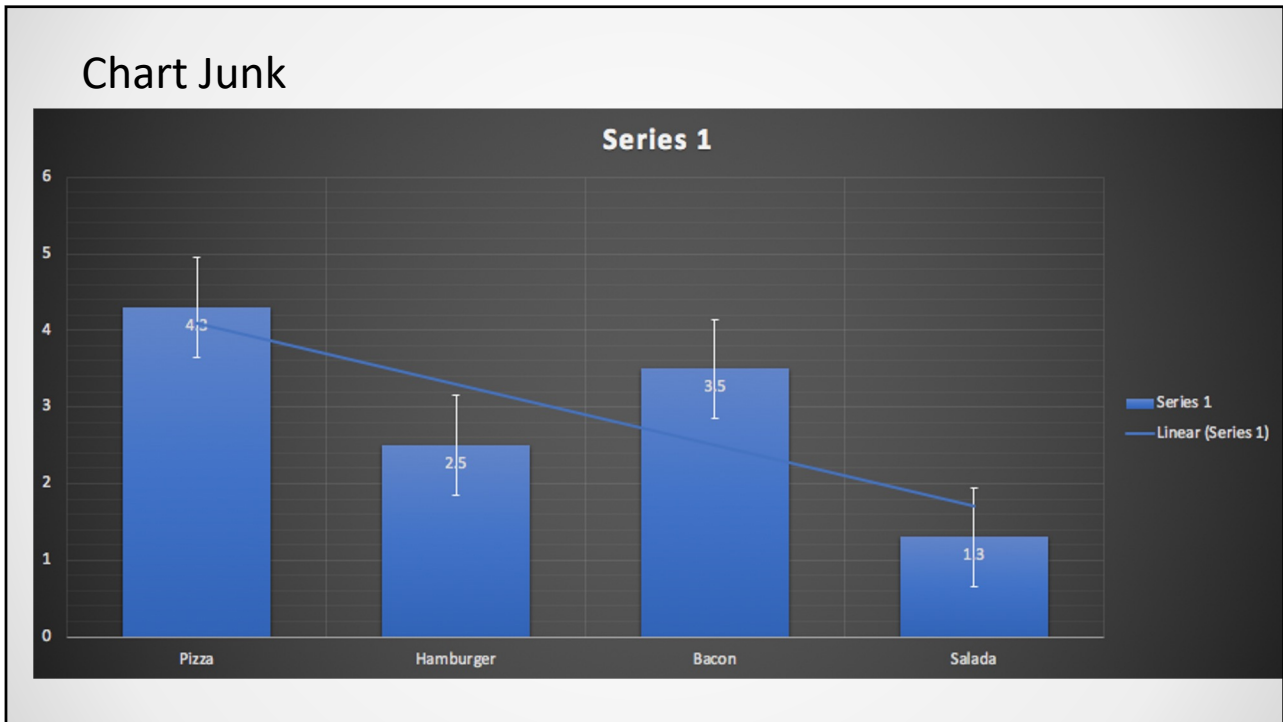
21

Chart Junk

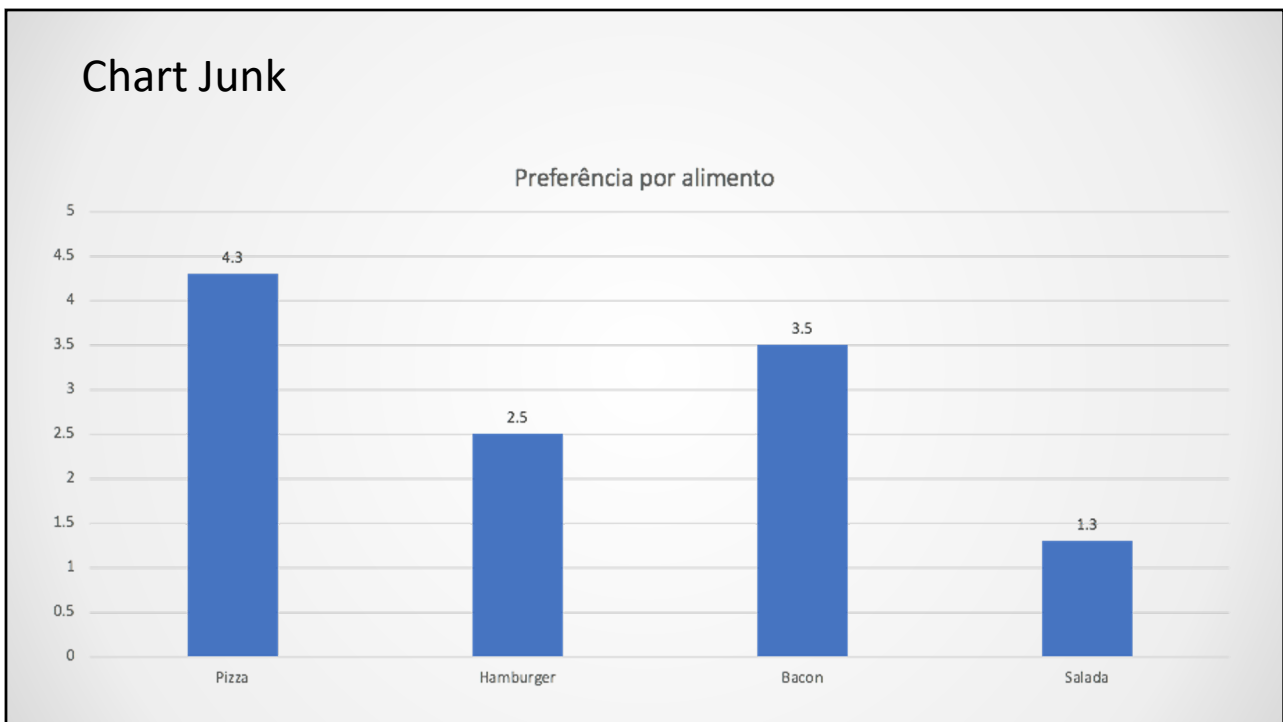
- When deciding which components we should use, it is a good idea to think about what **NOT** to use
- Tufte created the "data-ink ratio" concept, which is the relation between the ink required to plot the data and the ink used for the rest



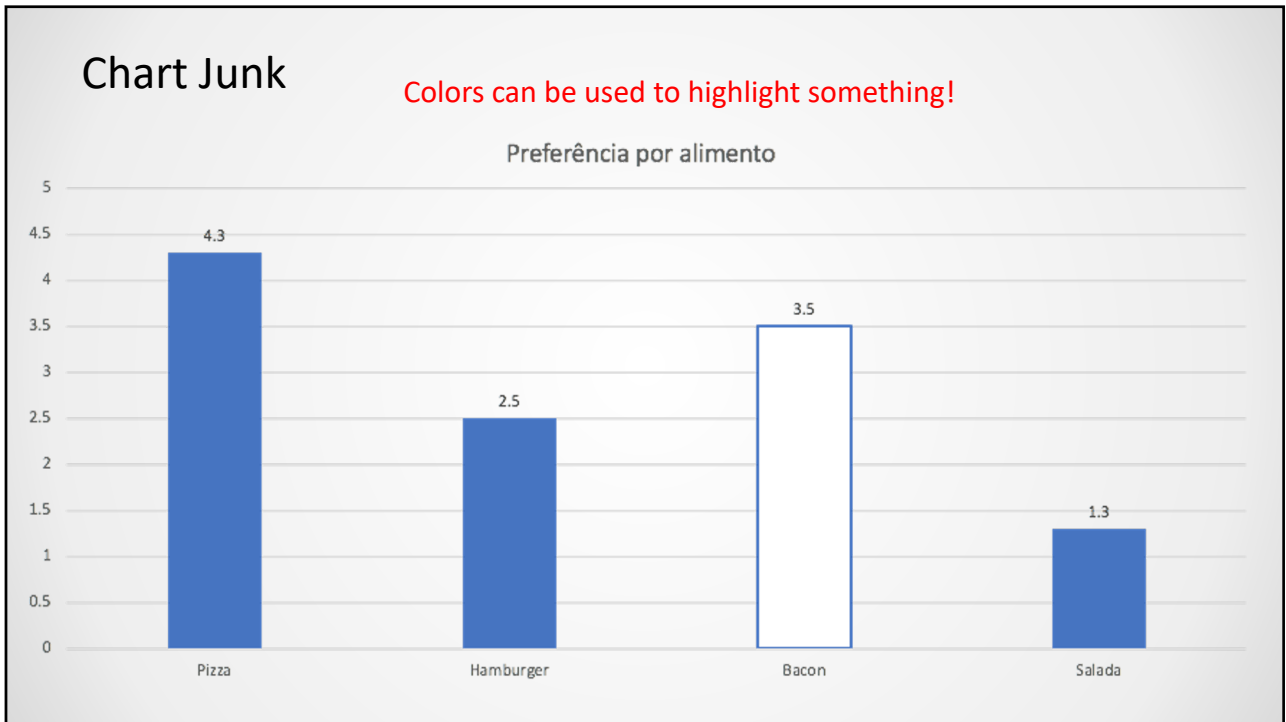
22



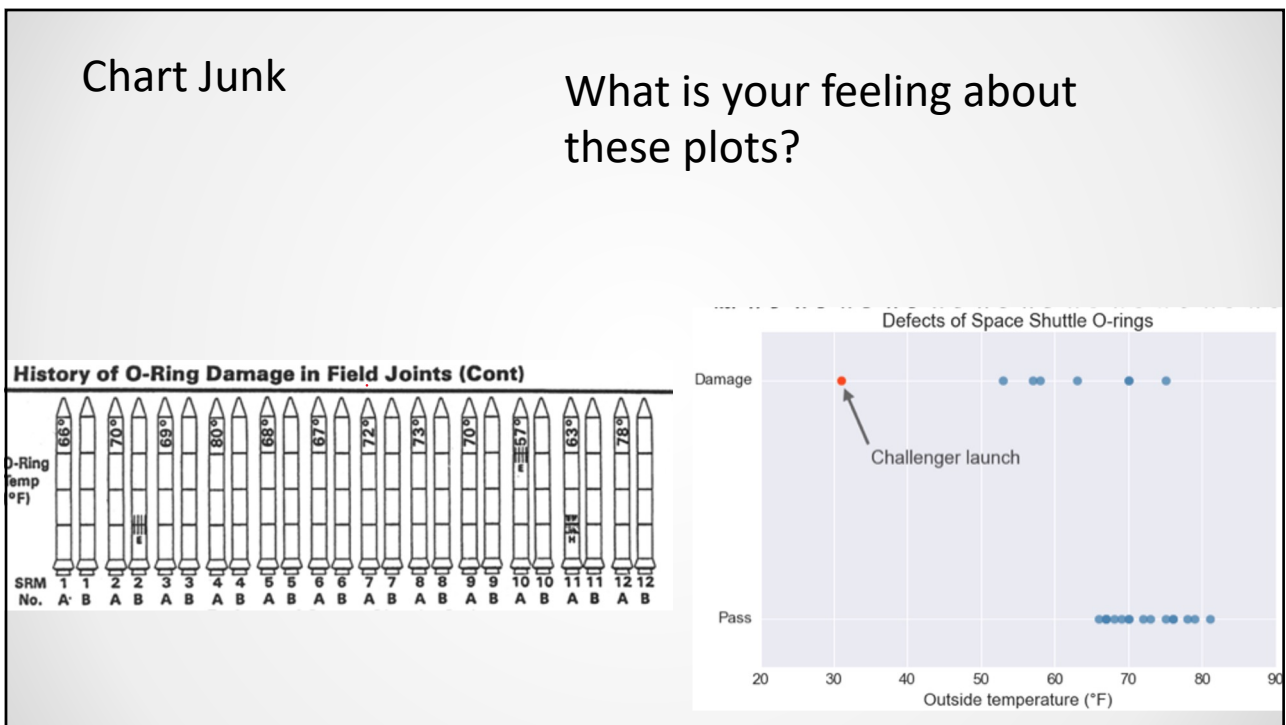
23



24



25

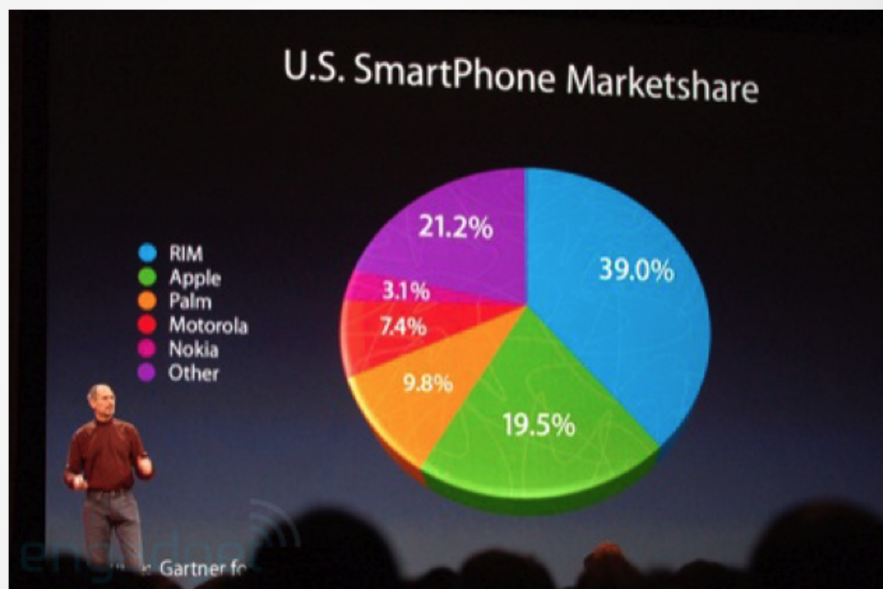


26

ANALYSIS

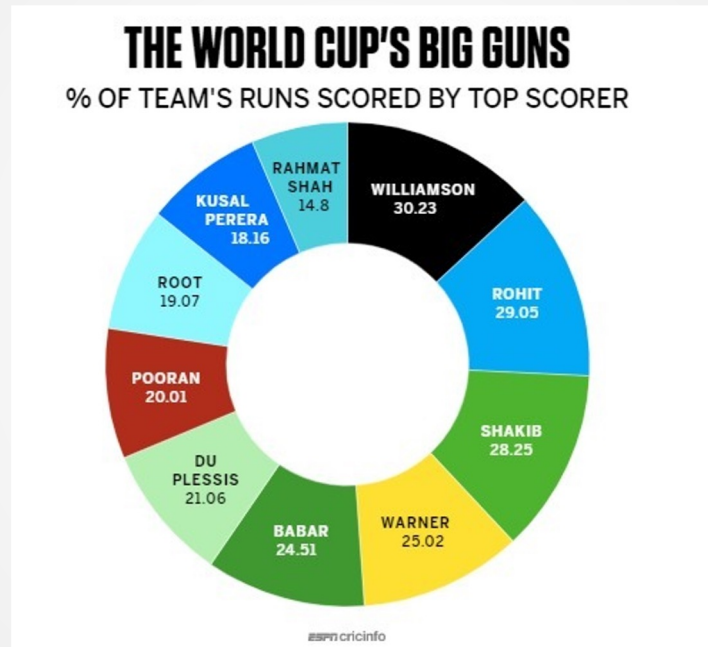
27

What problems do you see here?



28

And here?



29

Mortes de Covid-19 no Brasil

75% das vítimas tinham doenças associadas

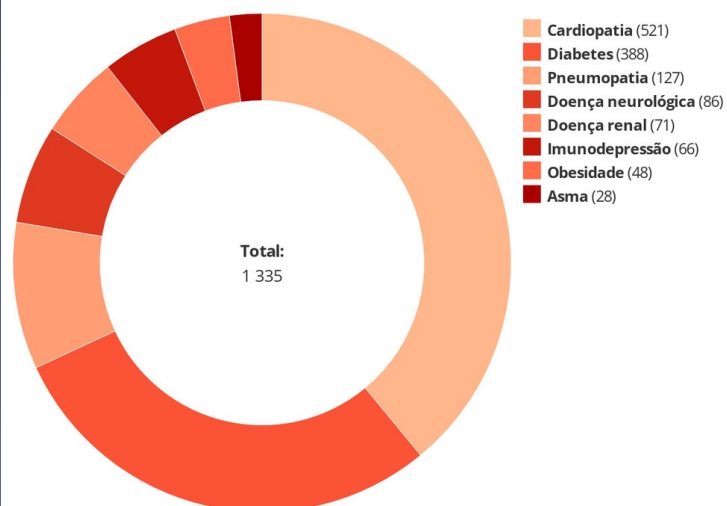
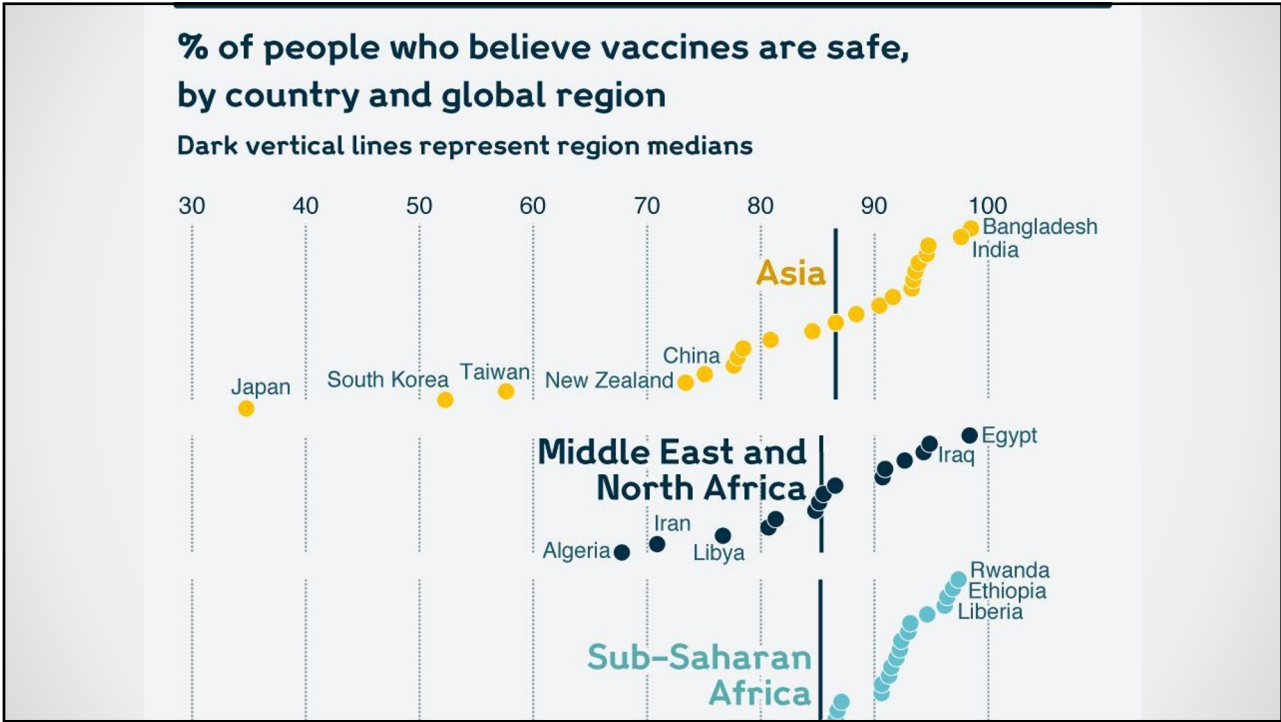
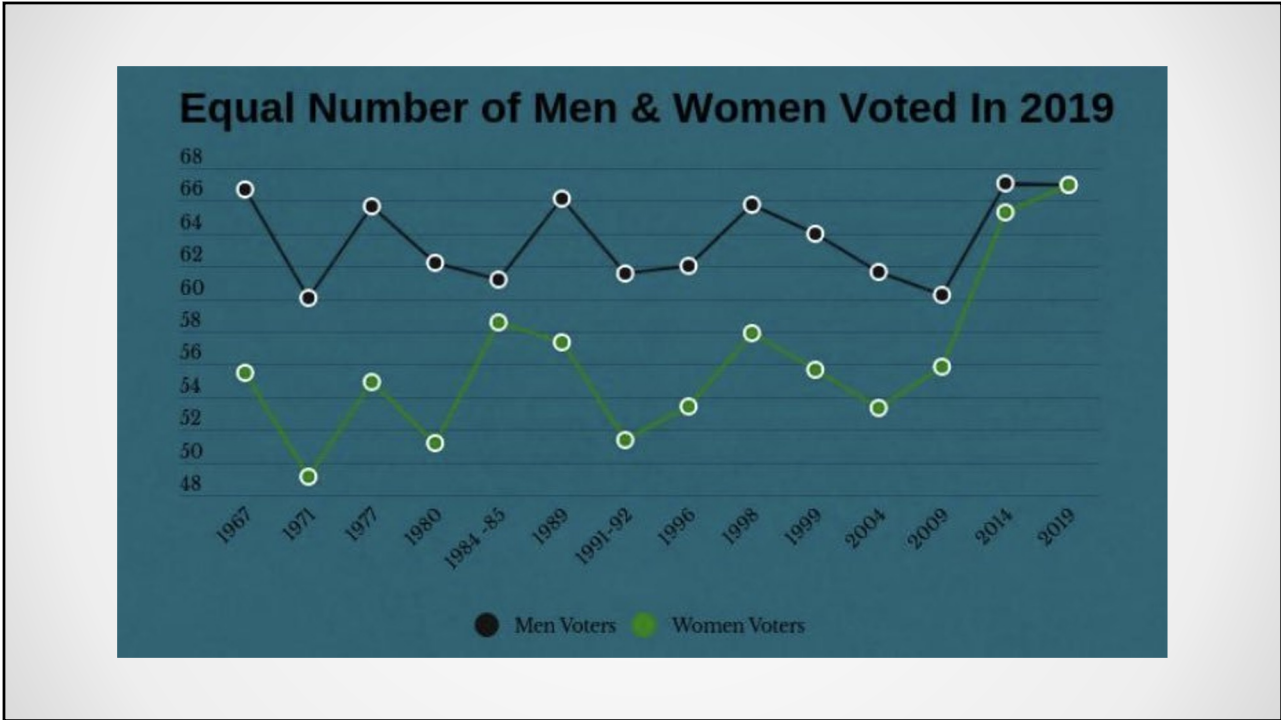


Gráfico: G1 • Fonte: Ministério da Saúde

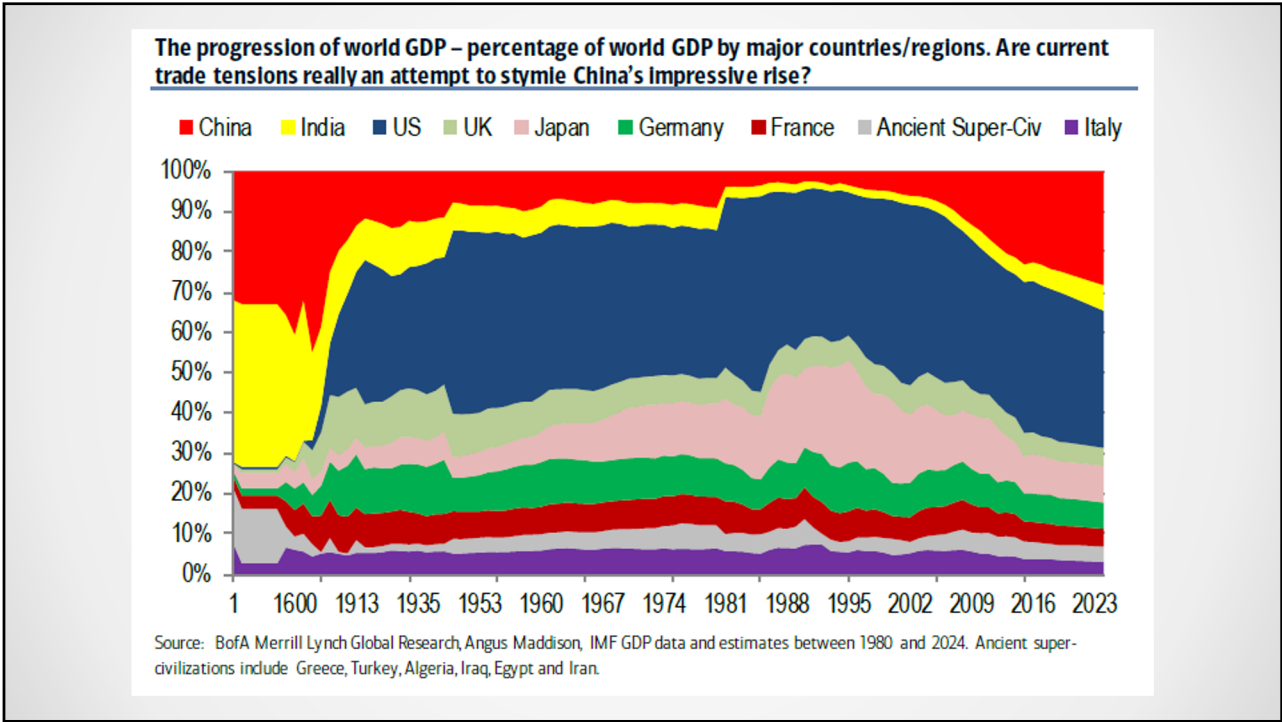
30



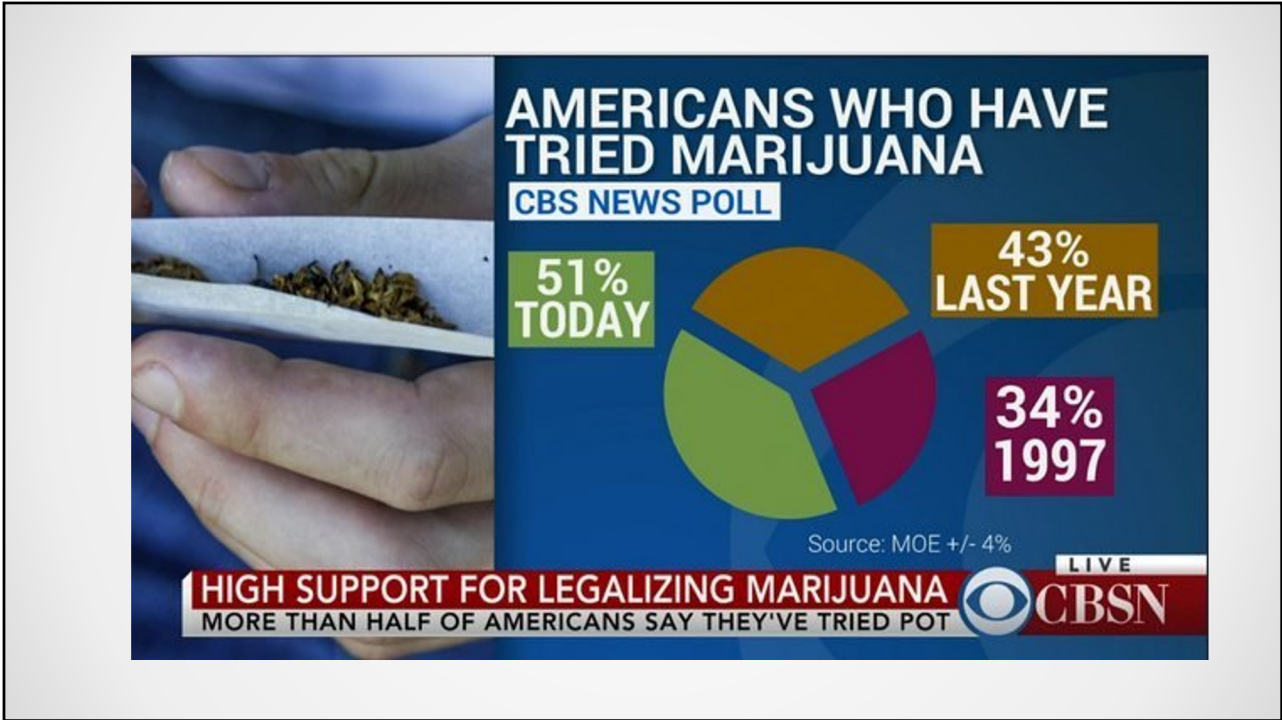
31



32



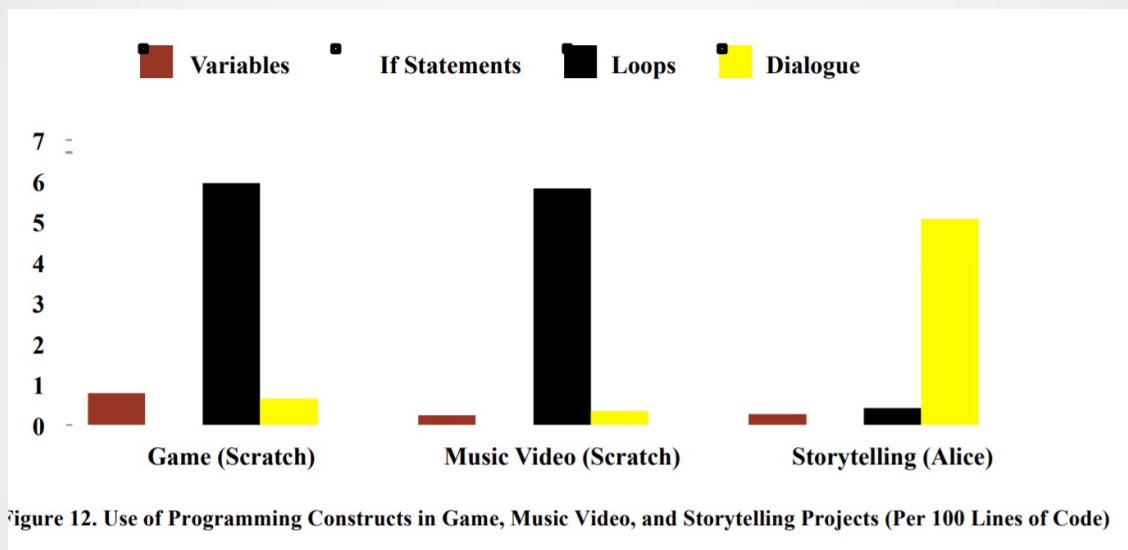
33



34

If you're not satisfied enough...

35



36

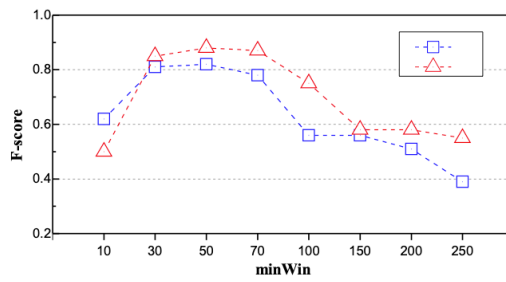


Fig. 14. F-scores obtained with different minWin

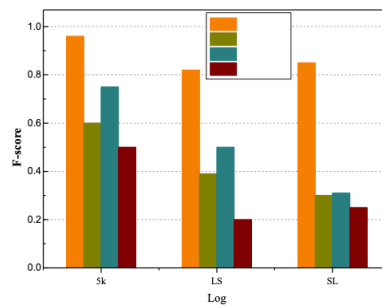
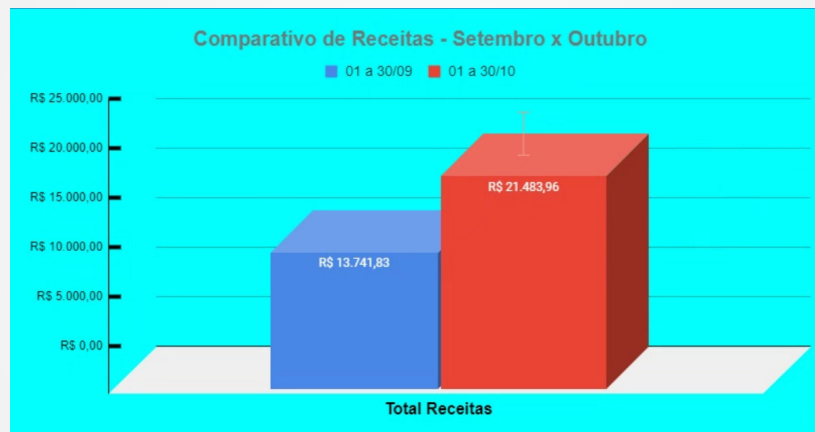


Fig. 15. F-scores obtained under different adaptive window strategies

37



38

References

Most of these visualizations were obtained from
<https://badvisualisations.tumblr.com/>

39

GESTALT PRINCIPLES

40

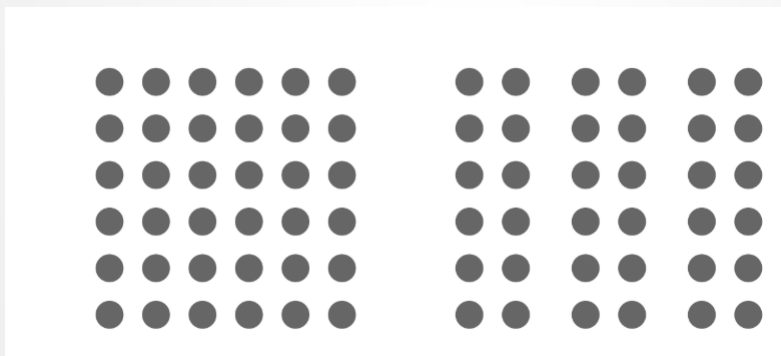
Gestalt principles

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

41

Proximity

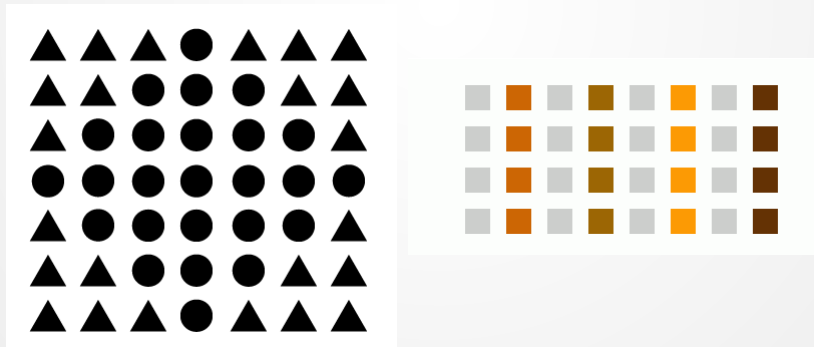
- Things that are closer to one another are perceived as belonging to a same group



42

Similarity

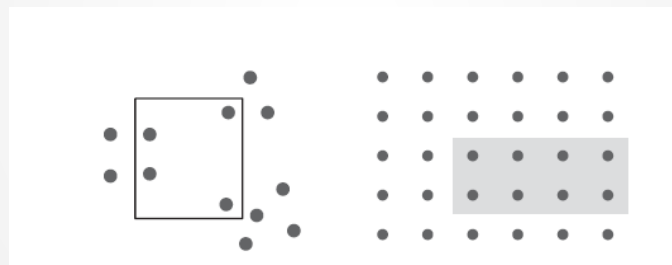
- Objects that share shapes, sizes, colors or orientation are perceived as belonging to the same group



43

Enclosure

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



44

Closure

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



45

Continuity

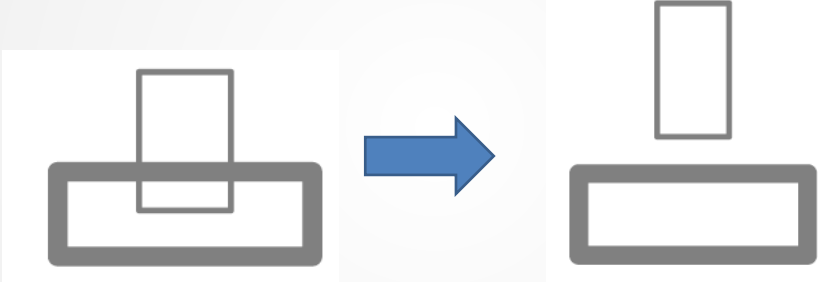
- Our eyes seek the most “natural” and “smooth” path between objects, even though they may not exist

How would you separate these items?



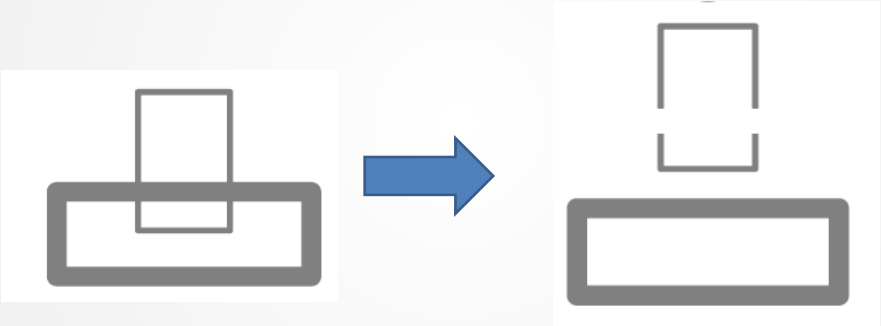
46

Continuity - 1



47

Continuity - 2



48

Connection

- We tend to perceive connected objects as a group

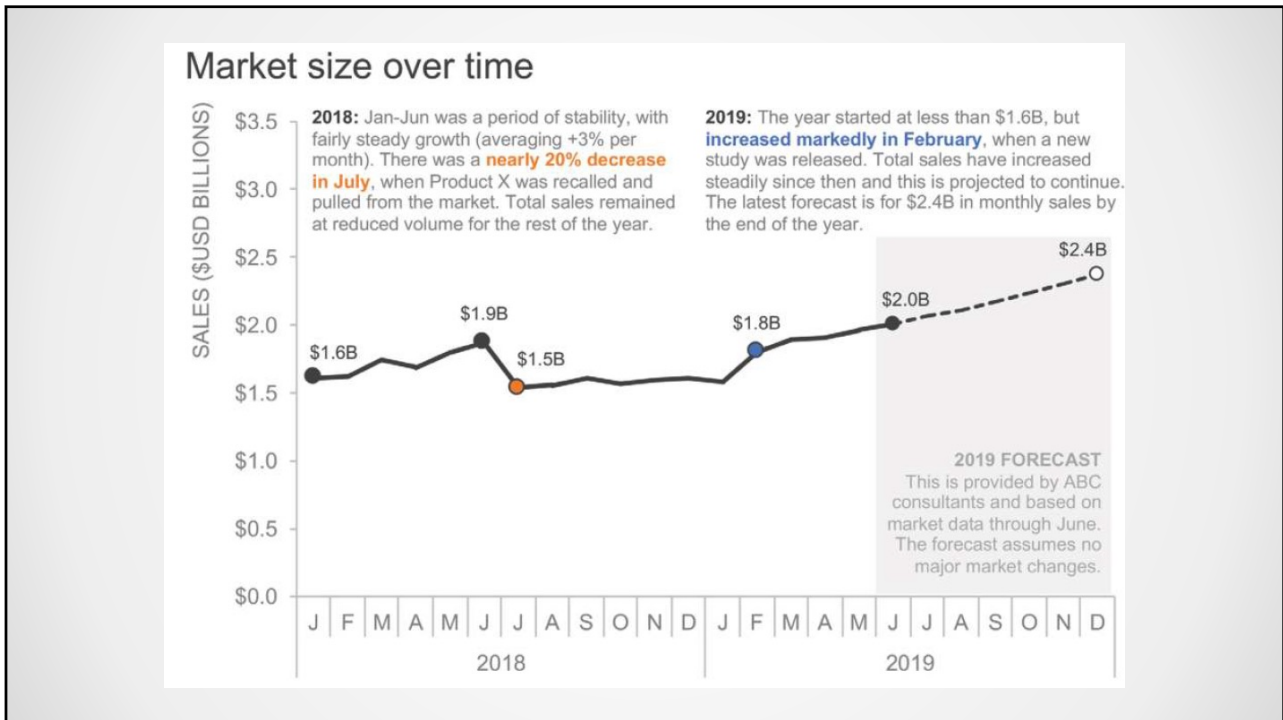


- Connection is often stronger than color, shape or size

49

Analyze the following visualization (5 minutes).
Next, you will be asked a (tough) question about it and
the gestalt principles.

50



51

Which Gestalt principles have been used?

52

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

53

Which gestalt principles have been used?

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

54

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

55

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

56



57

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

58



59

2. Remove the grid lines

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

60



61

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

62



63

4. Eliminate diagonal texts in the x axis

- Diagonal and vertical texts are polemic
- Whenever possible, prefer horizontal texts

64

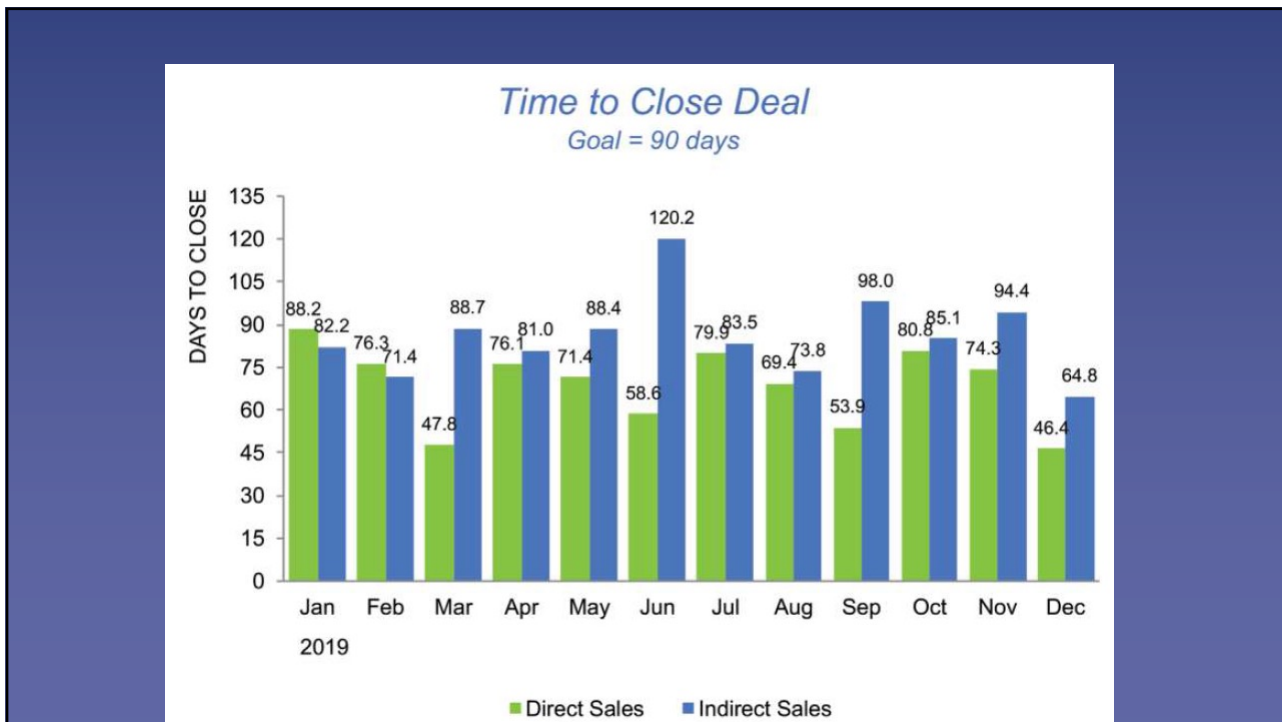


65

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

66

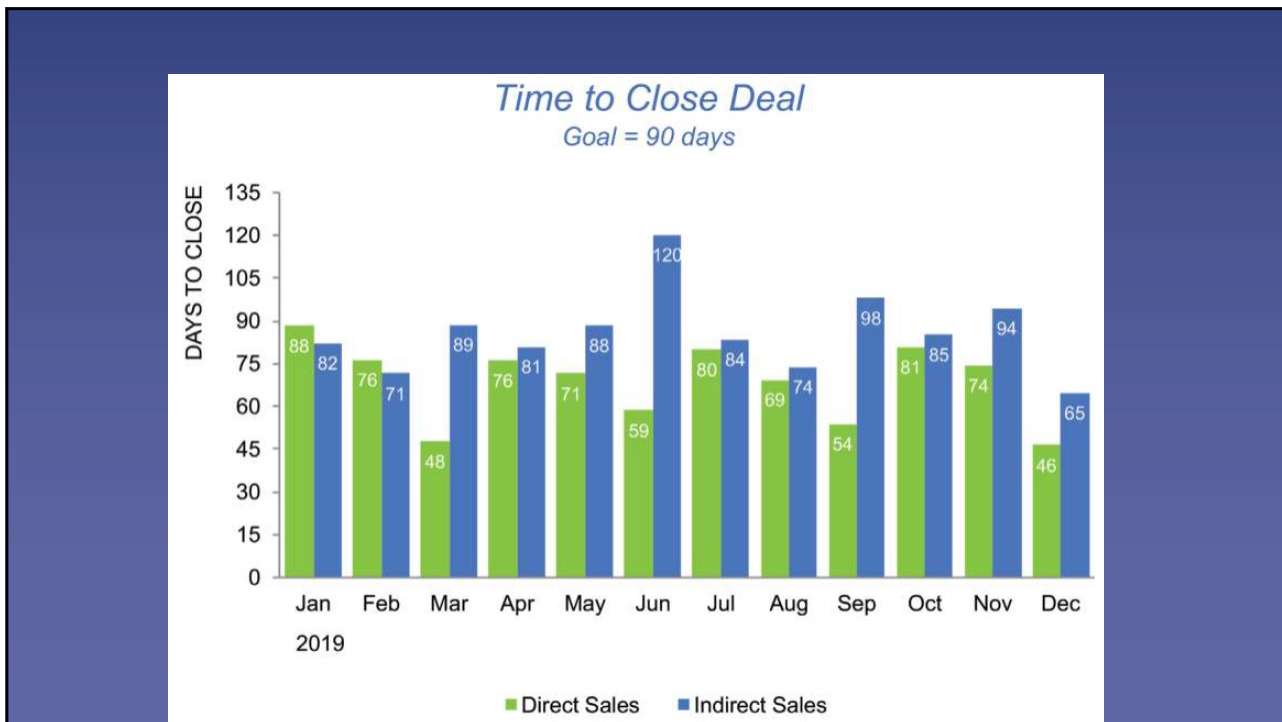


67

6. “Drag” the labels to the bars

- Whenever possible, round the values

68

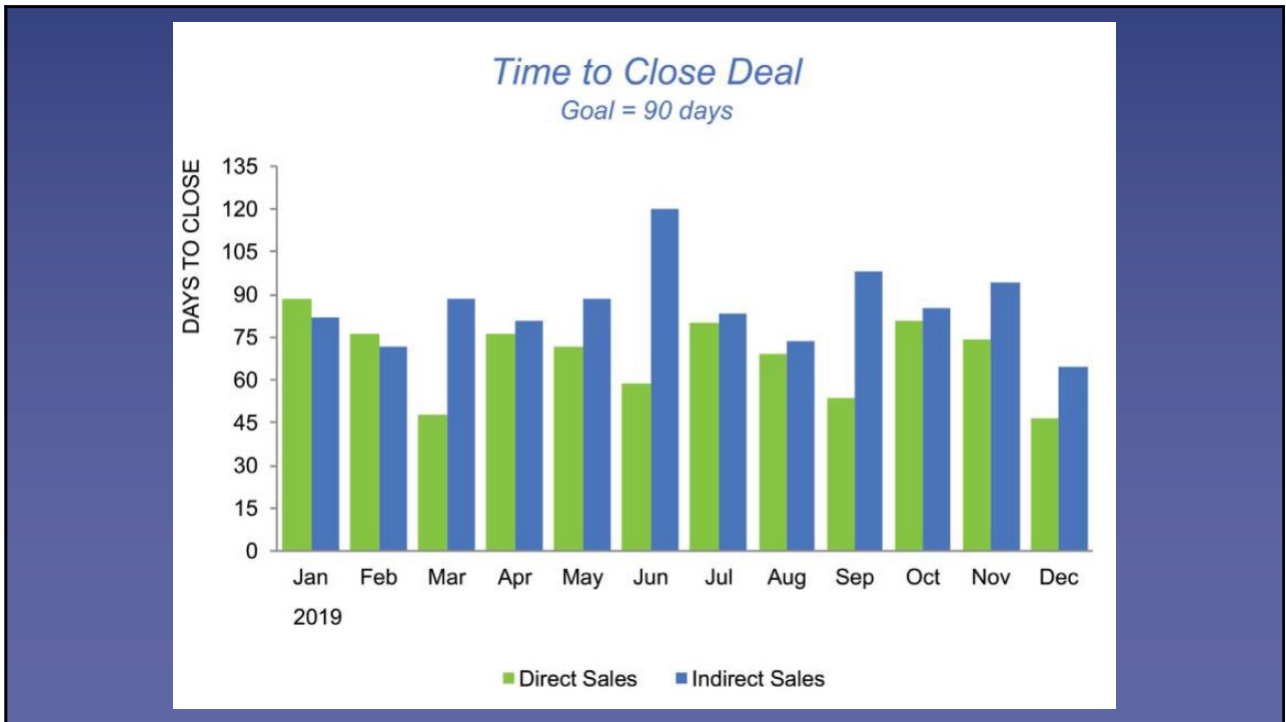


69

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?

70



71

8. Make it a line plot

72



73

9. Apply the legend to the plot

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

74

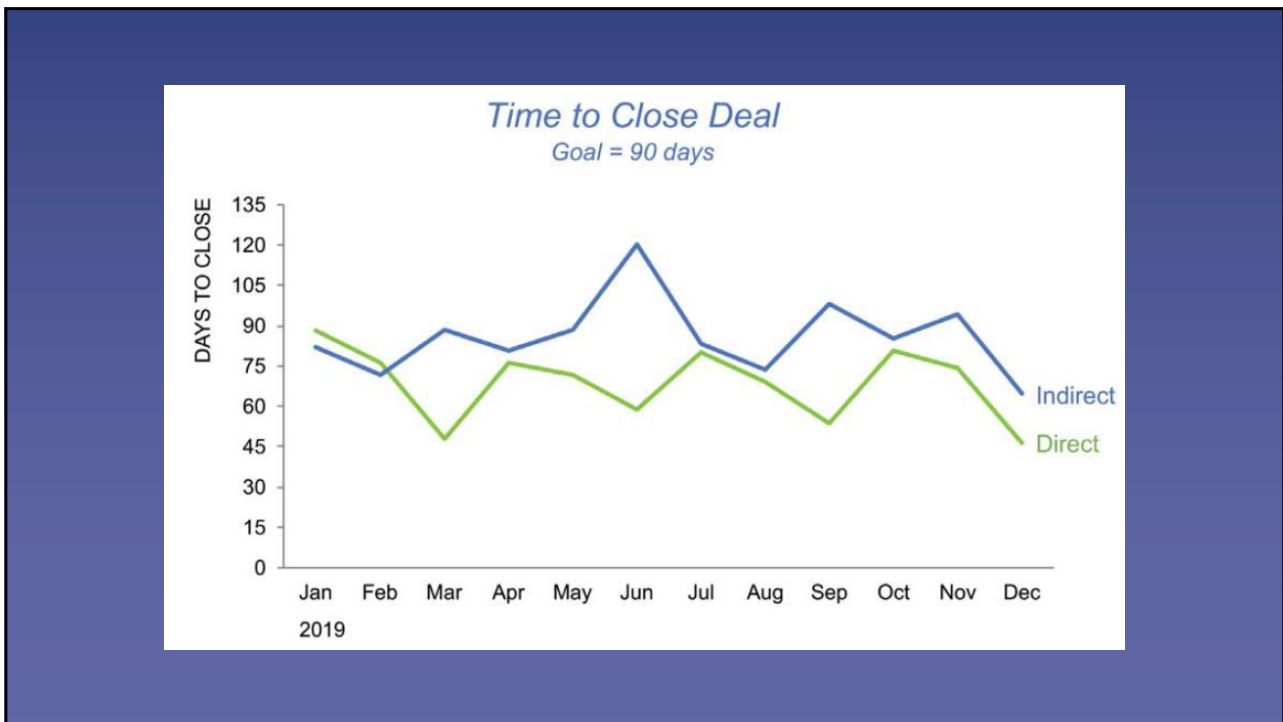


75

10. Changing the legend color to adhere to the data

- Proximity and similarity

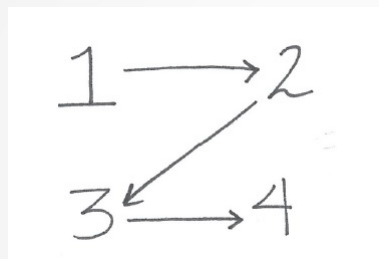
76



77

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

78

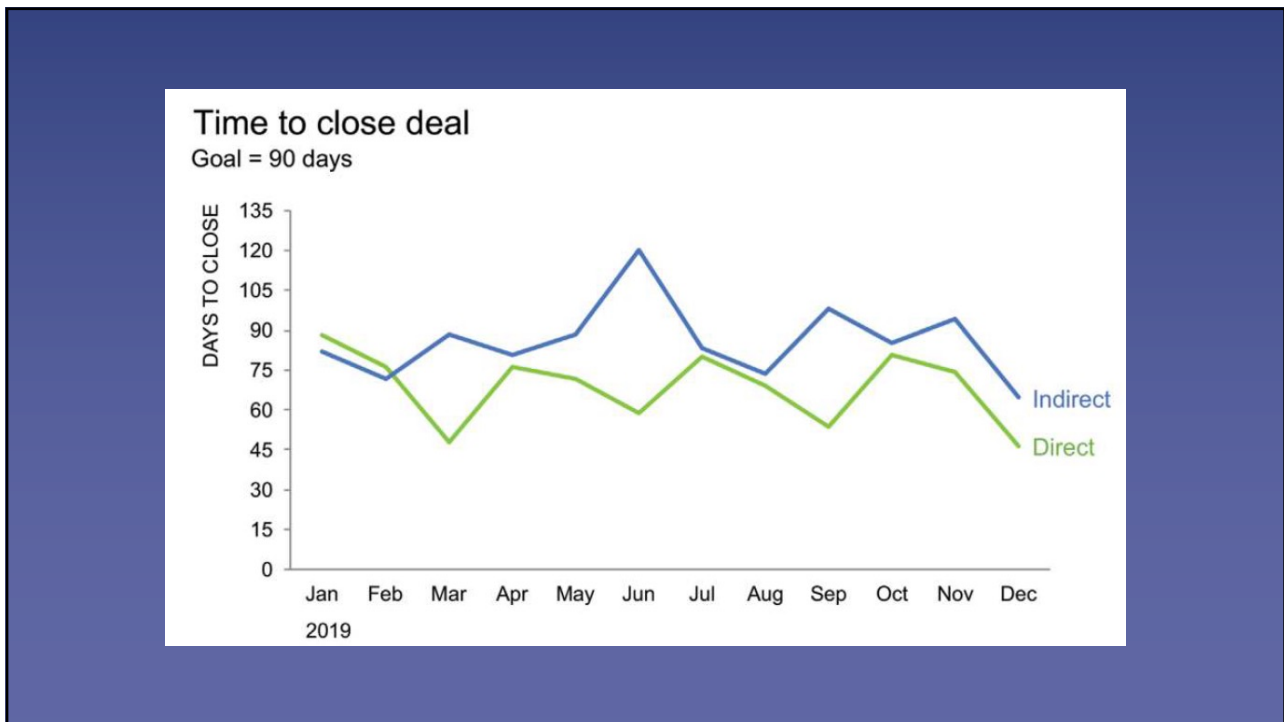


79

12. Removing the title color

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

80

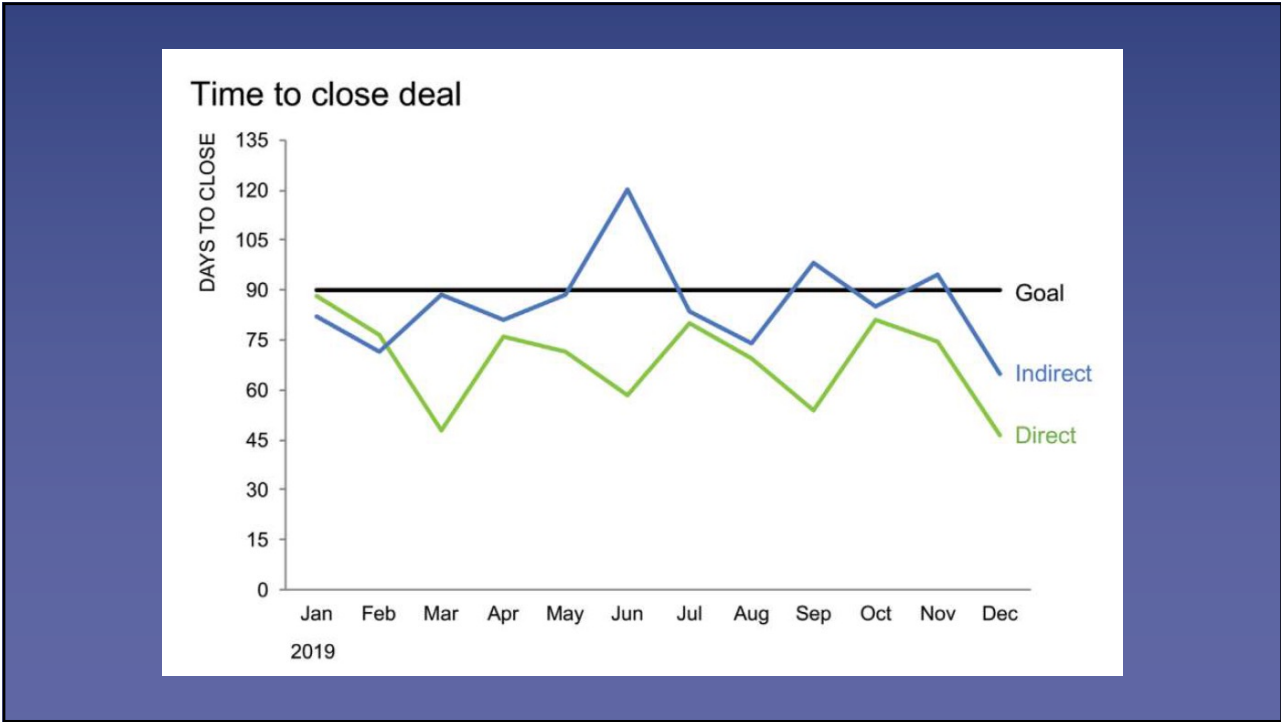


81

13. Adding the goal to the plot

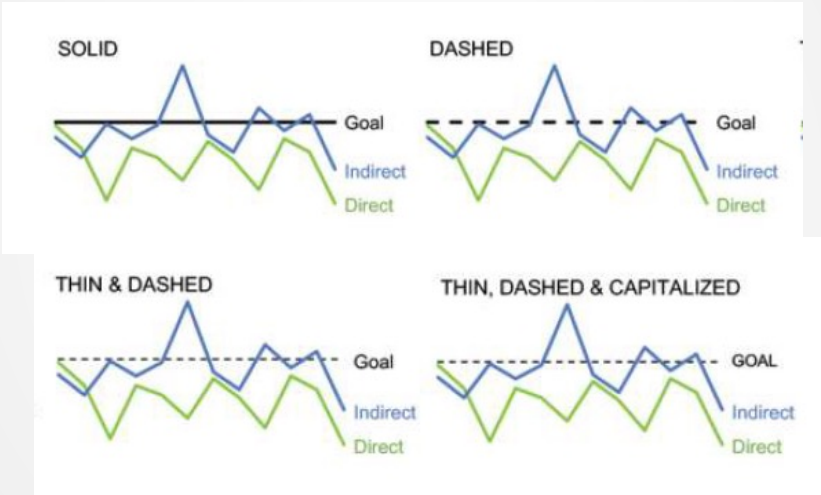
- The goal to make a deal is 90 days
- This information can be added to the plot to make the analysis of the curves more visual

82

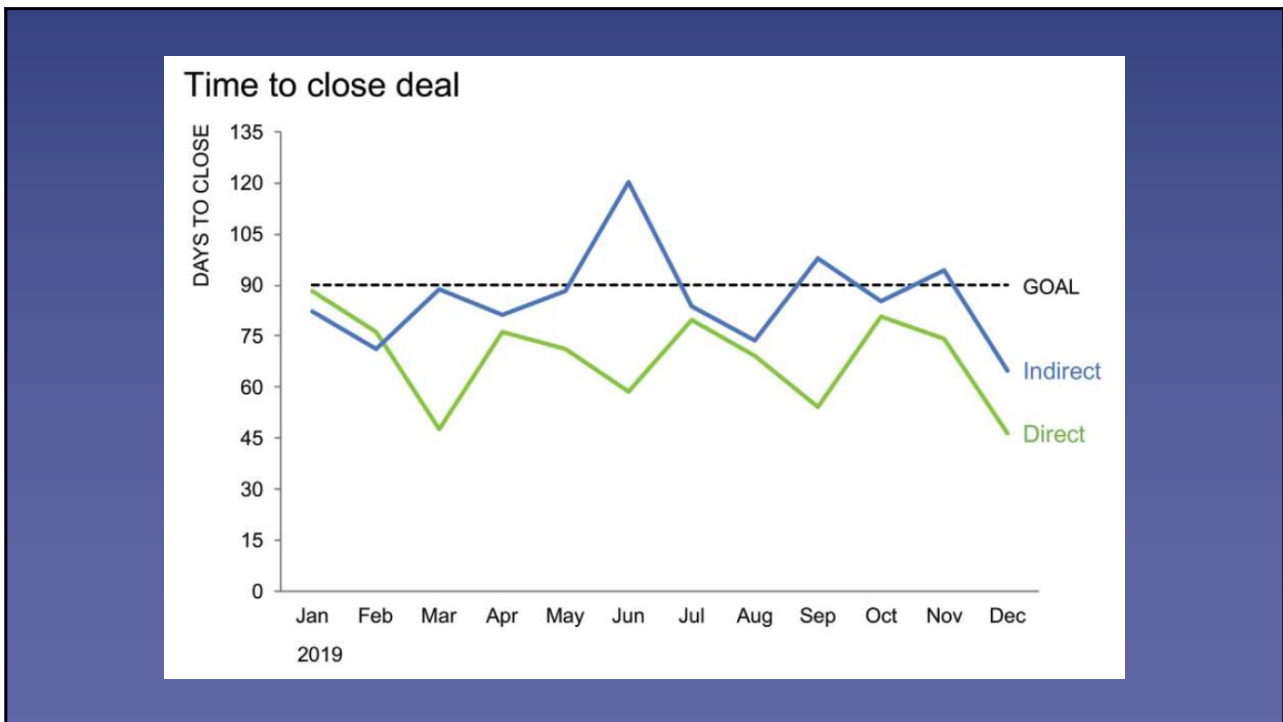


83

14. Testing different approaches



84

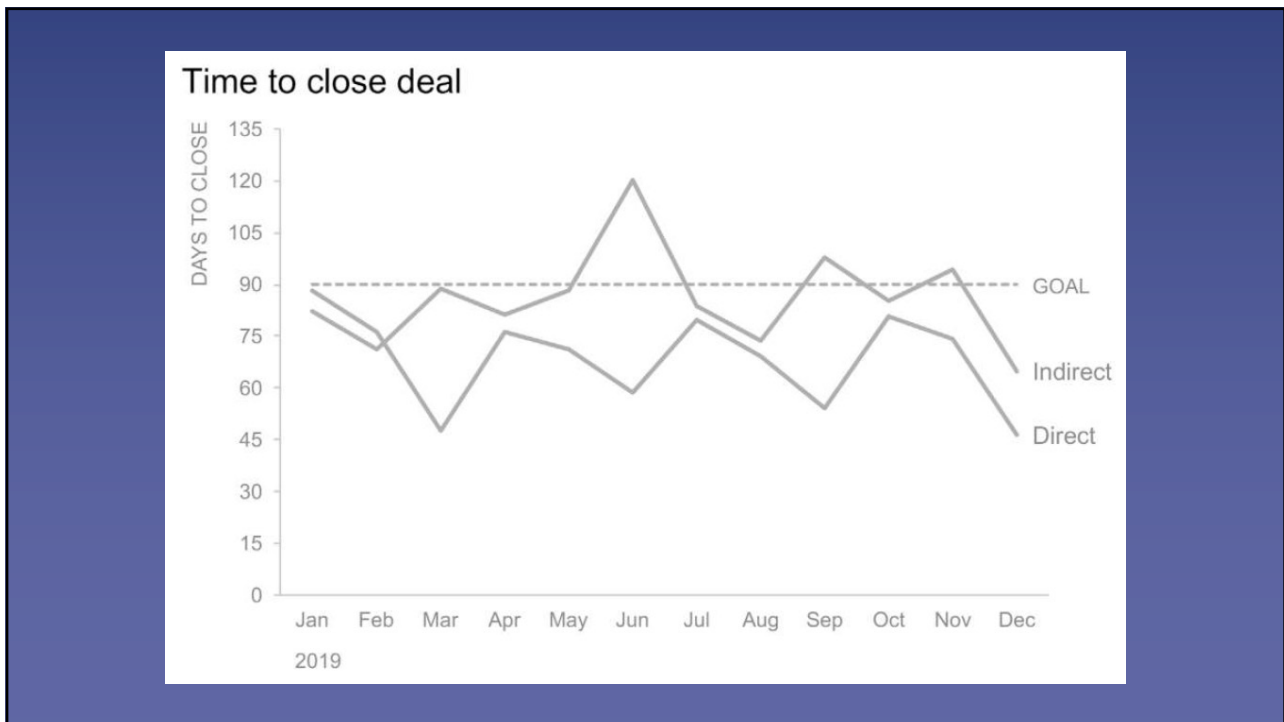


85

15. Removing colors

- We have enough separation between the lines

86

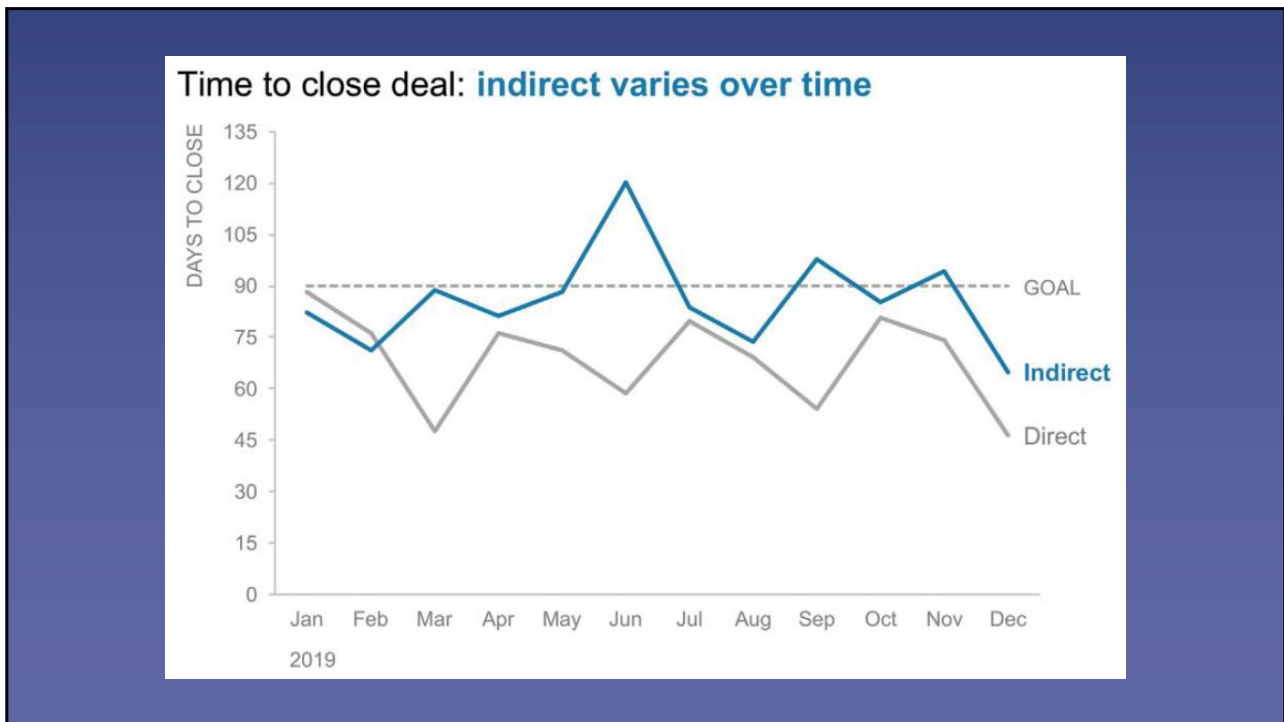


87

16. Drawing attention

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

88



89

17. Focusing on other aspects

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

90



91

BE MINDFUL OF COLORS

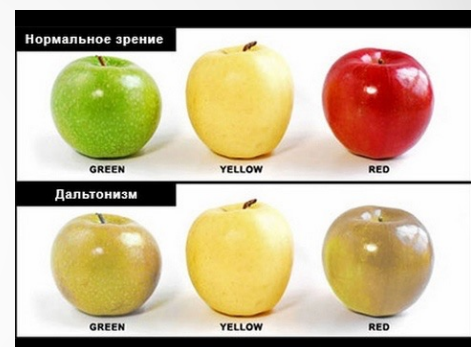
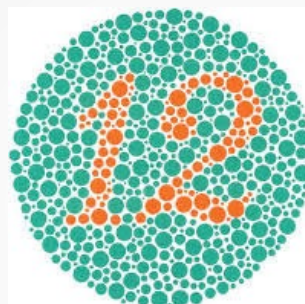
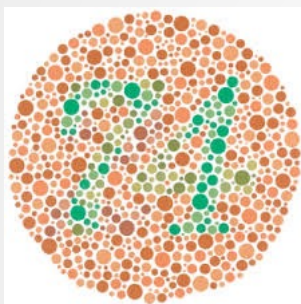
92

Colors

- One of the most common 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

93

Color blindness



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

94

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>

95

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**.

96

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

97

References



98