

ISA 401: Business Intelligence & Data Visualization

14: Fundamentals of Data Visualization

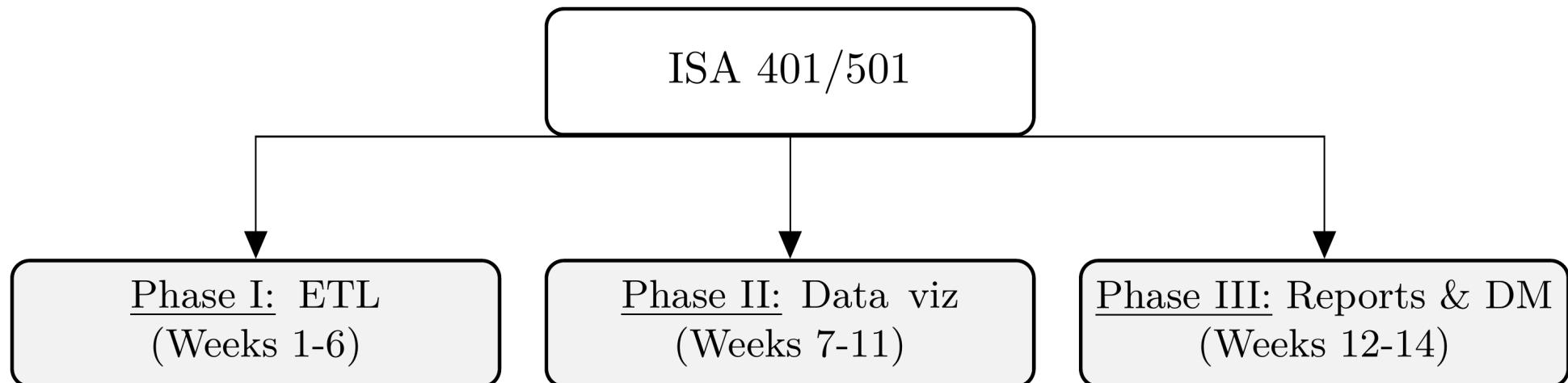
Fadel M. Megahed, PhD

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University

Twitter: [FadelMegahed](#)
GitHub: [fmegahed](#)
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Virtual Office Hours

Spring 2022

Refresher: Organization of this Course



How the ISA 401/501 course is organized.

Learning Objectives for Today's Class

- Explain the concept of "graphical excellence"
- Explain the theory of data graphics
- Optimize visual encoding based on data types
- Understand why color should be used sparingly and how to select appropriate colors (when color is a must)

Graphical Excellence

05 : 00

Non-graded activity: Terrible Charts

Activity

Russia's Defense Budget

White House Economy Growth

Tucker Carlson

Over the next 5 minutes, please identify the **1-2 main problems** in the charts in the following tabs.

- Write down your answers in the editable area of each chart.
- Discuss your answers with your neighboring classmates.
- Be prepared to share these answers with class.

05 : 00

Non-graded activity: Terrible Charts

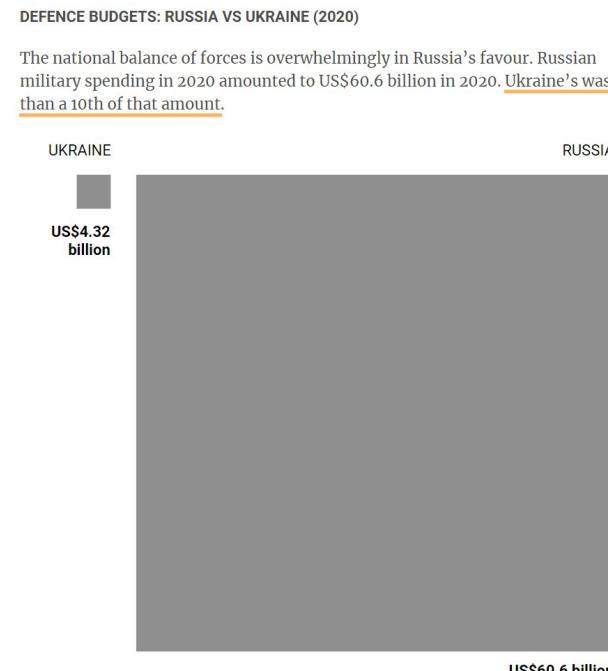
Activity

Russia's Defense Budget

White House Economy Growth

Tucker Carlson

Source: The chart was embedded in [this tweet](#) by Cedric Scherer; however, it is unclear which news outlet have created the original chart.



Main Issue(s): (Insert below)

Non-graded activity: Terrible Charts

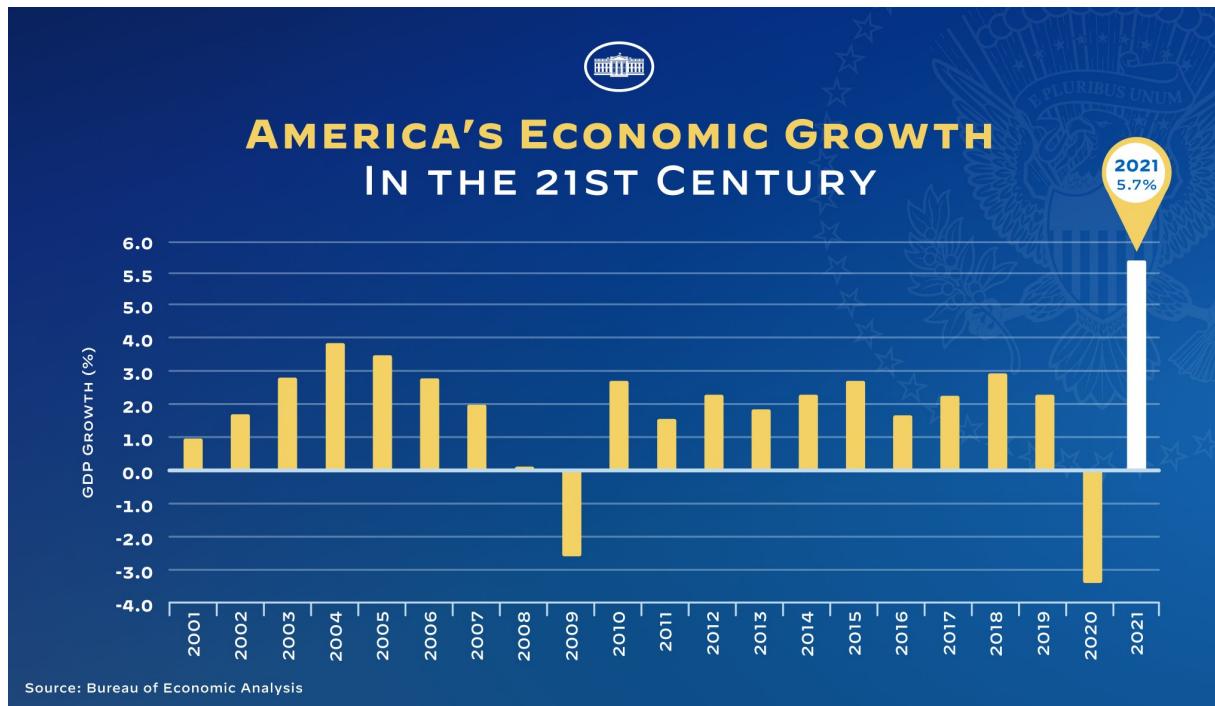
Activity

Russia's Defense Budget

White House Economy Growth

Tucker Carlson

Source: The chart was created by the White House and shared via [this tweet from the verified White House account](#). Note that the chart was latter correct.



Main Issue(s): (Insert below)

05 : 00

Non-graded activity: Terrible Charts

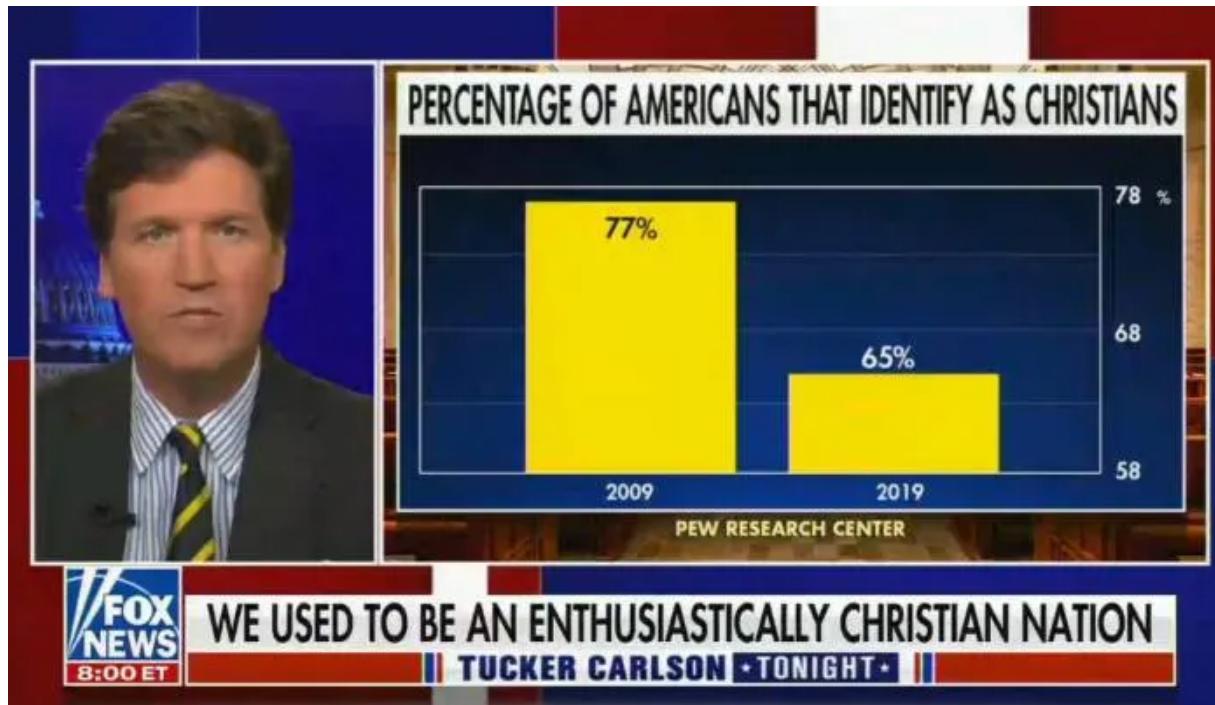
Activity

Russia's Defense Budget

White House Economy Growth

Tucker Carlson

Source: The chart was created by Fox and was highlighted in [this Fox News Clip](#).

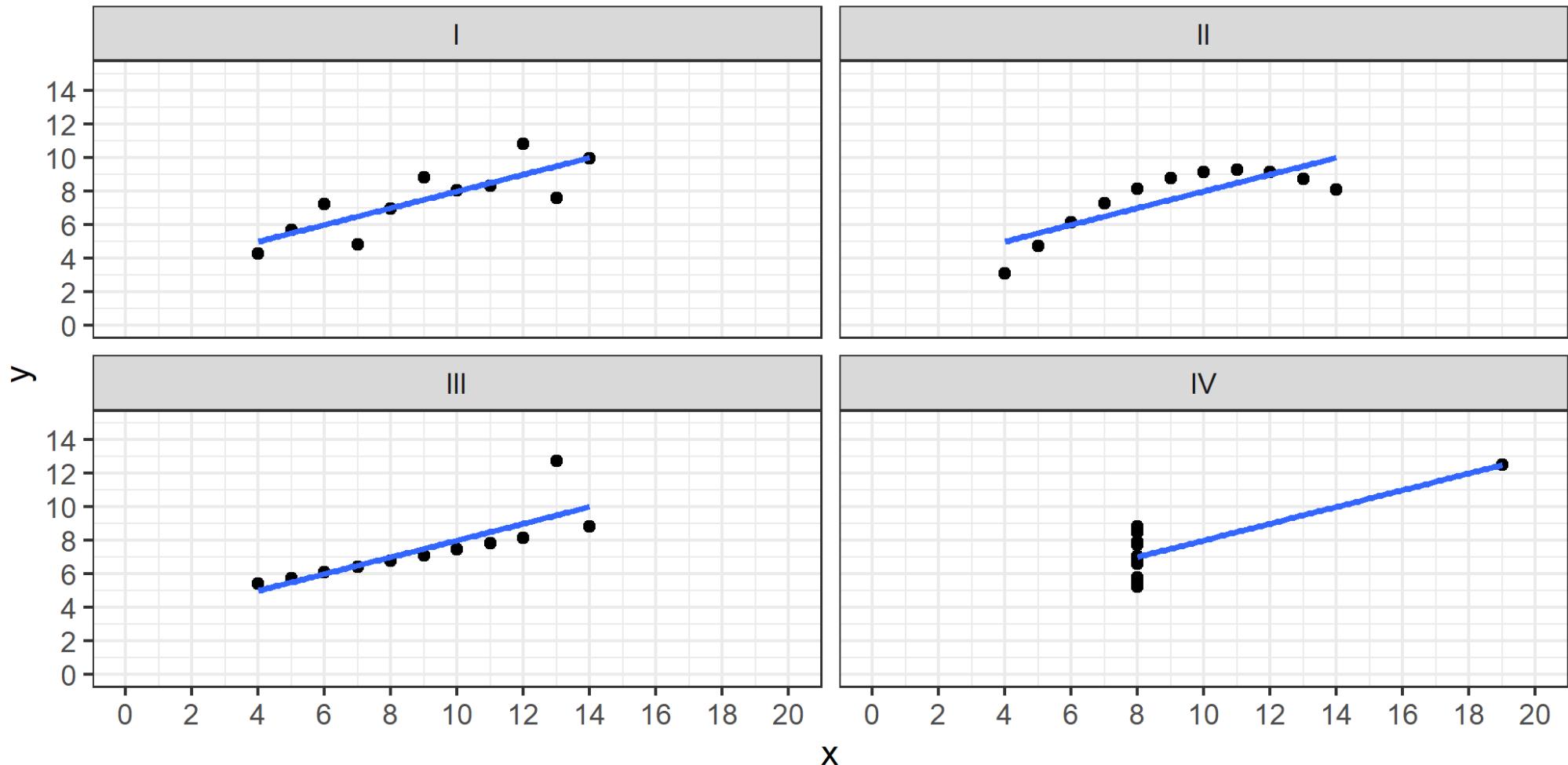


Main Issue(s): (Insert below)

Graphical Excellence: What should Graphs Do?

- Show the data
- Lead to thinking about the **substance** rather than something else
- Avoid **distorting** what the data have to say
- Present **many numbers in a small space**
- Make **large datasets coherent**
- Encourage the eye to **compare different pieces of the data**
- **Reveal the data at several levels of detail**, from a broad overview to the fine structure
- Serve **a purpose**: description, exploration, tabulation, decoration
- Be **closely integrated with the statistical & verbal descriptions of the data**

Show/Reveal the Data: Anscombe's Dataset



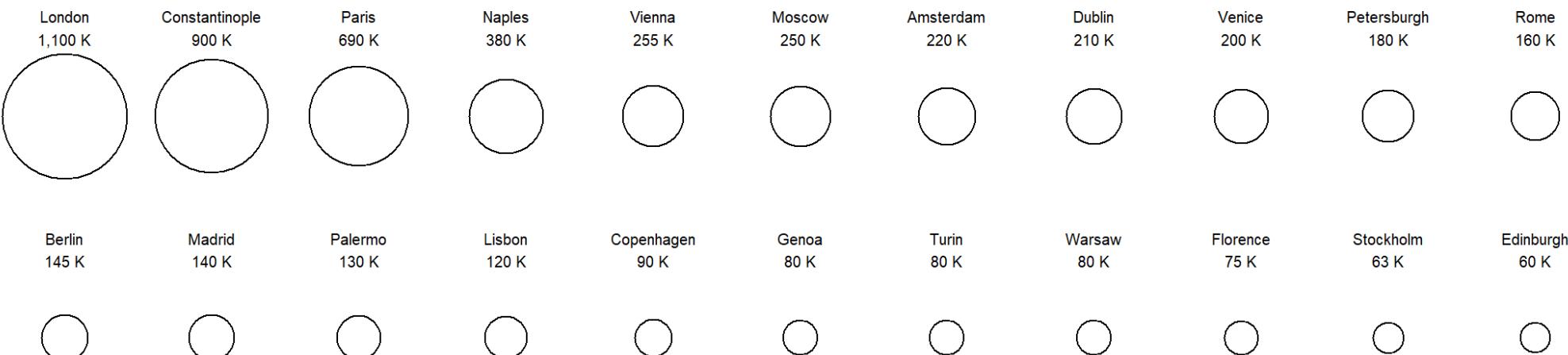
Substance

Activity

Your Solution

My Solution

In 5 minutes, please **sketch** a better (non-bubble) chart than the one use by William Playfair for plotting the populations of 22 European cities at the end of the 1700s.



Substance

Activity

Your Solution

My Solution

Ideally, on a piece of paper sketch out your solution. Otherwise, please feel free to download the plot's data (using the code below) and use a software of your choice for plotting a better chart for the data.

```
pacman:::p_load(tidyverse)

playfair = read.table("http://www.stat.uiowa.edu/~luke/data/Playfair") %>%
  rownames_to_column(var = 'city') %>% # converting row names to city var
  as_tibble() %>% # converting it to a tibble
  arrange( desc(population) ) # arranging the rows in a descending order by population

write_csv(x = playfair, file = 'playfair_data.csv')
```

Be prepared to share your solution with the entire class.

05 : 00

Substance

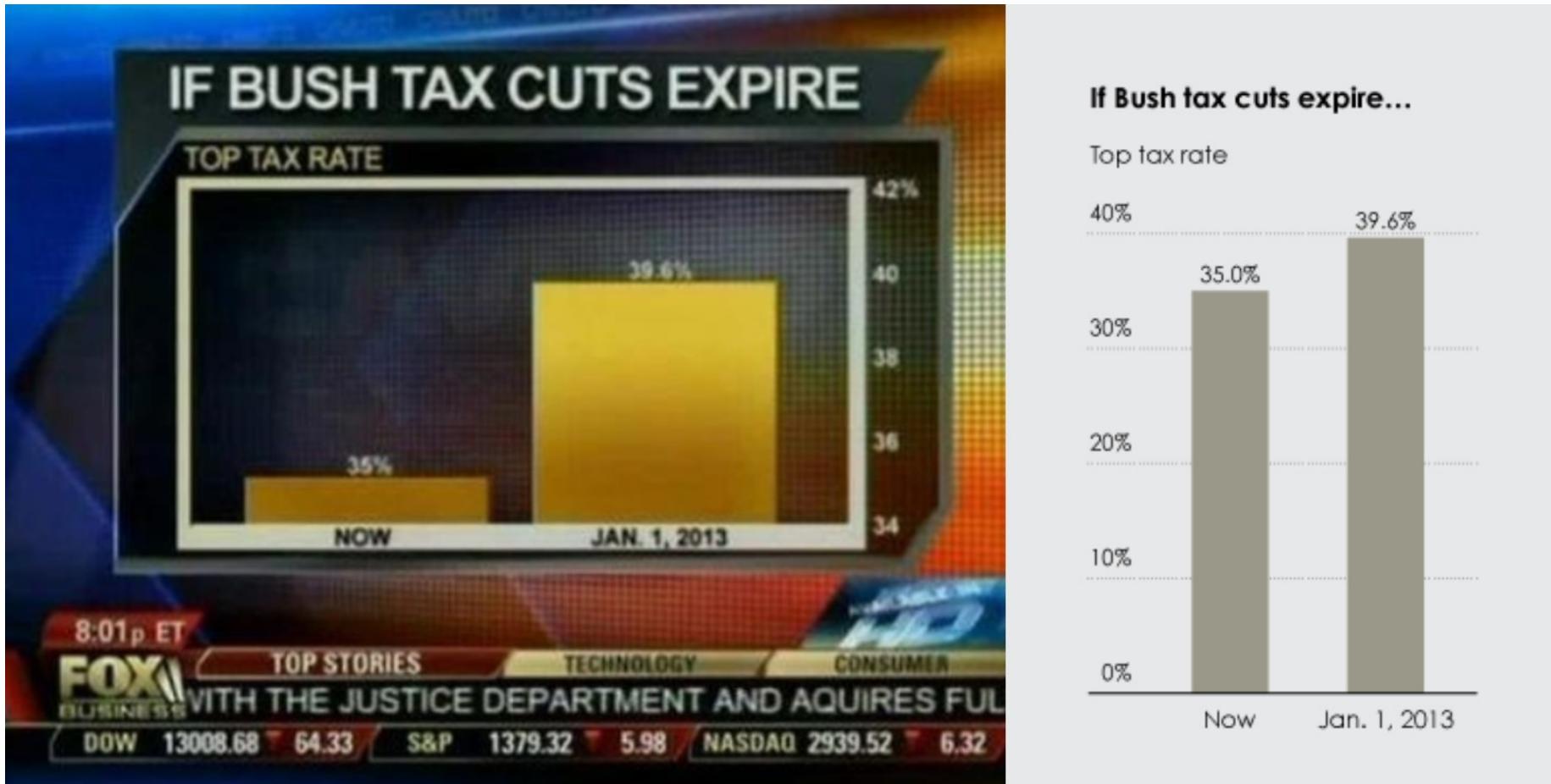
Activity

Your Solution

My Solution

- In my opinion, a **dot chart is more effective** than the bubble chart. The **population would be mainly encoded using the position**; you can still use area as a secondary encoding mechanism.

Avoid Distortion of Data

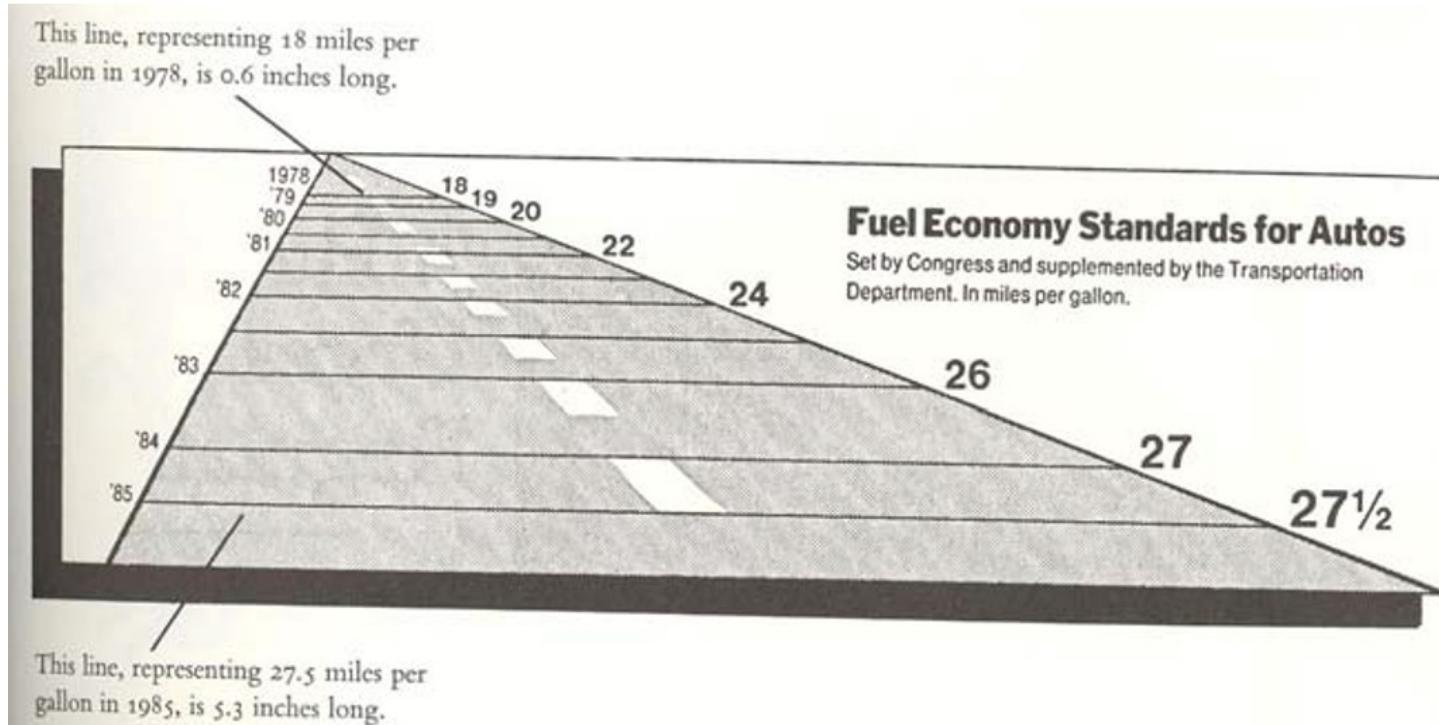


Source: Nathan Yau (2012). Fox News continues charting excellence.

Avoid Distortion of Data: The Lie Factor

Size of effect shown in graphic

Size of effect in data



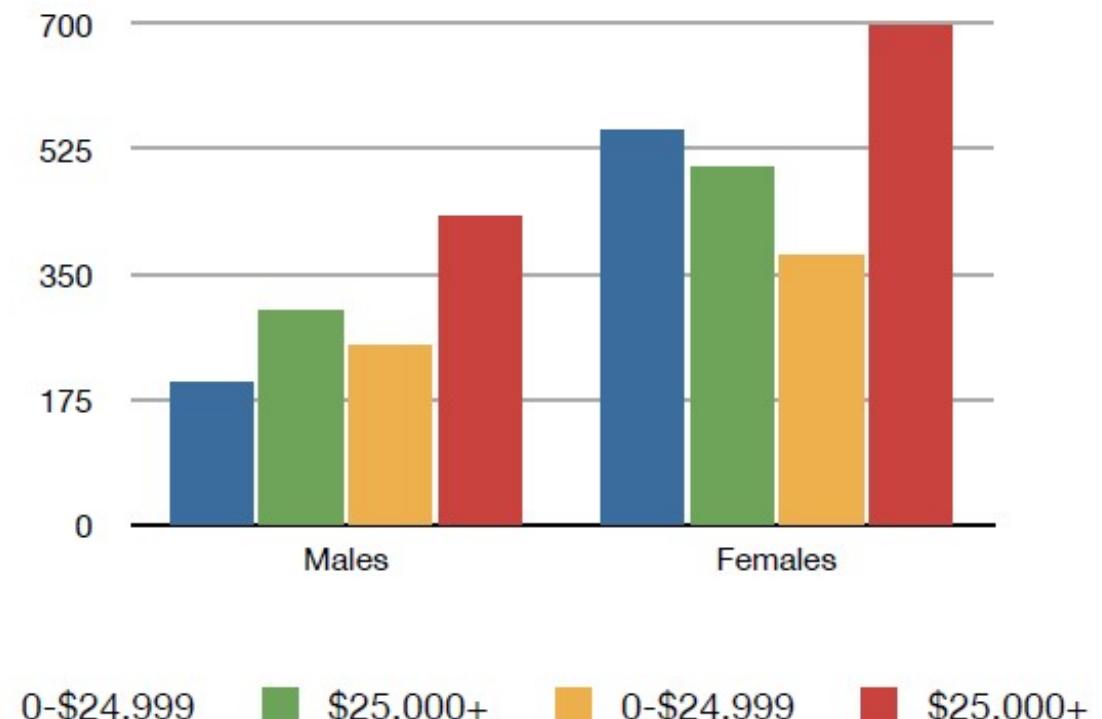
Graphical Integrity Principles: A Summary

- Clear, detailed, and thorough labeling and appropriate scales
- Size of the graphic effect should be directly proportional to the numerical quantities (“lie factor”)
- Show data variation, not design variation

Theory of Data Graphics

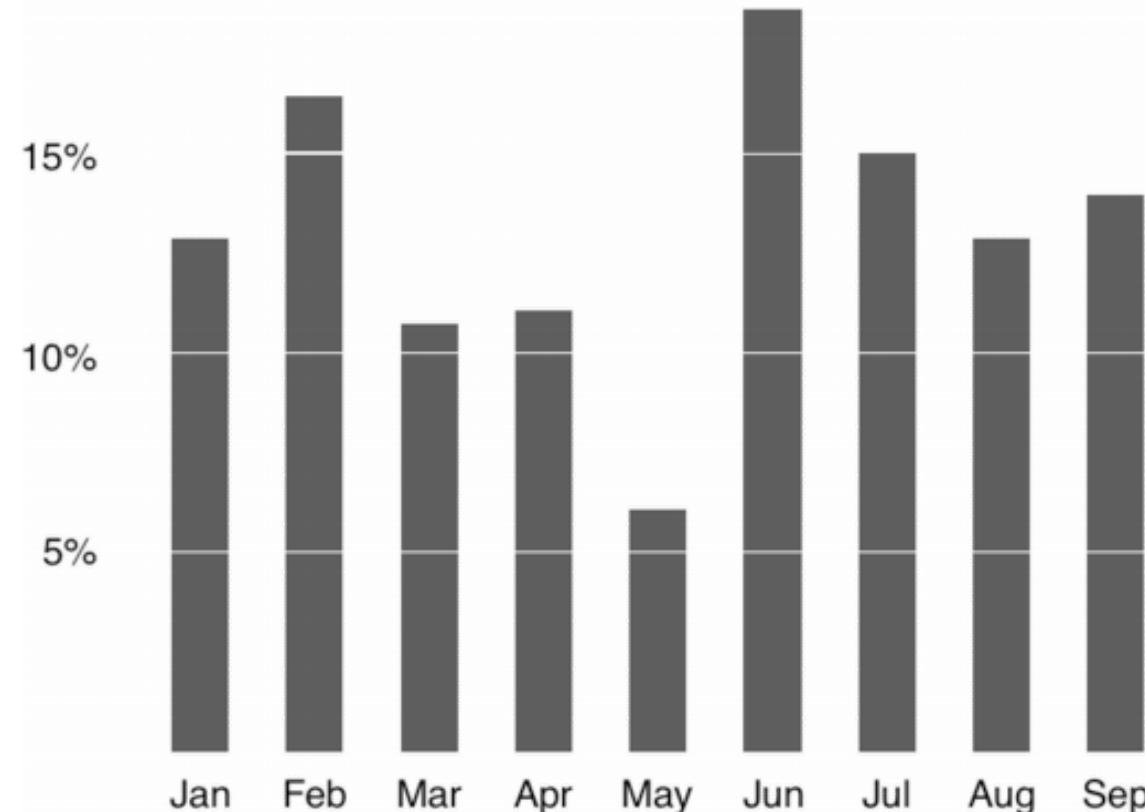
Definition of Data Ink

- Data-ink refers to the non-erasable ink used for presenting the data.
 - If data-ink would be removed from the image, the graphic would lose the content.



Focus on Data: Avoid Chartjunk

Chartjunk is the extraneous visual elements that distract from the message!!



Other Subjective Design Principles

- **Aesthetics:** Attractive things are perceived as more useful than unattractive ones
- **Style:** Communicates brand, process, who the designer is
- **Playfulness:** Encourages experimentation and exploration
- **Vividness:** Can make a visualization more memorable

Data Models

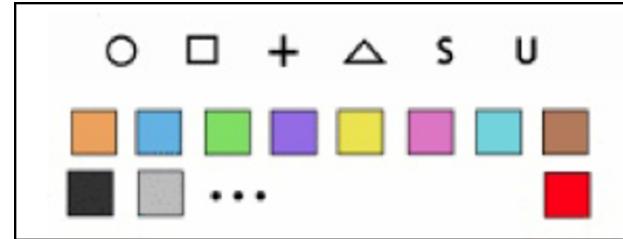
Data Types (from S. Stevens, Theory of Scales)

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	<i>Permutation group</i> $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	<i>Isotonic group</i> $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	<i>General linear group</i> $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	<i>Similarity group</i> $x' = ax$	Coefficient of variation

Data Types: Explained

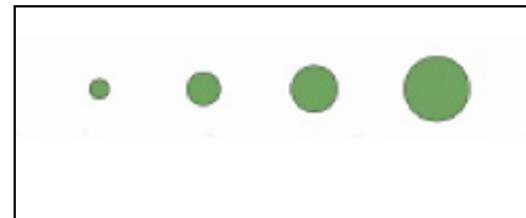
Nominal:

- Are = or \neq to other values
- Apples, bananas, oranges, etc.



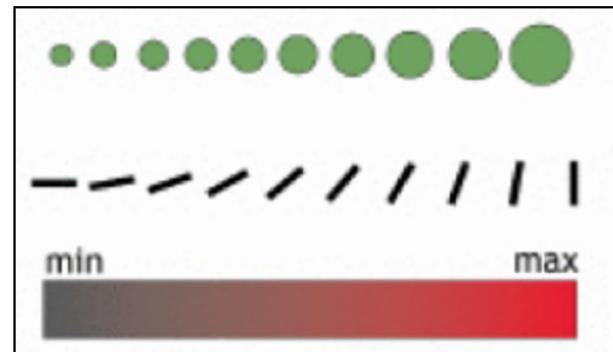
Ordinal:

- Obey a $<$ relationship
- Small, medium and large



Quantitative:

- Can do math on them
- 50 inches, 53 inches, etc.



Quantitative Data Types (from S. Stevens)

Quantitative data can be further divided into:

Intervals (Location of Zero Arbitrary):

- Dates: Jan 19; Location: (Lat, Long)
- Only differences (i.e., intervals) can be compared.

Ratio (Zero Fixed):

- Measurements: Length, Weight, Temperature, ...
- Origin is meaningful, we can compute ratios, proportions, differences, etc.

02 : 00

Non-Graded Activity: Data Terminology

Activity

Your Solution

In 2 minutes, please identify an appropriate data type for each column below.

Order ID	Order Date	Order Priority	Product Container	Product Cost	Ship Date
1	1/1/2022	5 - low	Large box	25	1/5/2022
2	1/4/2022	4 - not specified	Small Box	36	1/7/2022
3	1/15/2022	2- high	Small Box	38	1/17/2022
3	1/15/2022	2- high	Small Box	41	1/17/2022
3	1/15/2022	2- high	Jumbo Box	44	1/17/2022
3	1/15/2022	2- high	Wrap Bag	33	1/17/2022
4	1/18/2022	1- urgent	Small Box	33	1/19/2022

Showing 1 to 7 of 11 entries

Previous

1

2

Next

02 : 00

Non-Graded Activity: Data Terminology

Activity

Your Solution

Data Types: (Edit below)

- Order ID:
- Order/Ship Date:
- Order Priority:
- Product Container:
- Product Cost:

Data vs. Conceptual Models

From data model

- 32.5, 54.0, -17.3, ...

Using a conceptual model:

- Temperature

To data type:

- Continuous to x significant digits i.e. quantitative
- Hot, warm, cold i.e. ordinal
- Burned vs. not burned i.e. nominal

Image Model: Visual (Encoding) Variables

Channels

Position

Size

(Grey) Value

Texture

Color

Orientation

Shape

	Marks	Points	Lines	Areas
LES VARIABLES DE L'IMAGE				
XY 2 DIMENSIONS DU PLAN	POINTS	LIGNES	ZONES	
Z				
TAILLE				
VALEUR				
LES VARIABLES DE SÉPARATION DES IMAGES				
GRAIN				
COULEUR				
ORIENTATION				
FORME				

Mapping to Data Types

	Nominal	Ordinal	Quantitative
Position	✓	✓	✓
Size	✓	✓	~
(Grey)Value	✓	✓	~
Texture	✓	~	✗
Color	✓	✗	✗
Orientation	✓	✗	✗
Shape	✓	✗	✗

✓ = Good

~ = OK

✗ = Bad

Visual Channels and their Precision

Color Should Be Used Sparingly

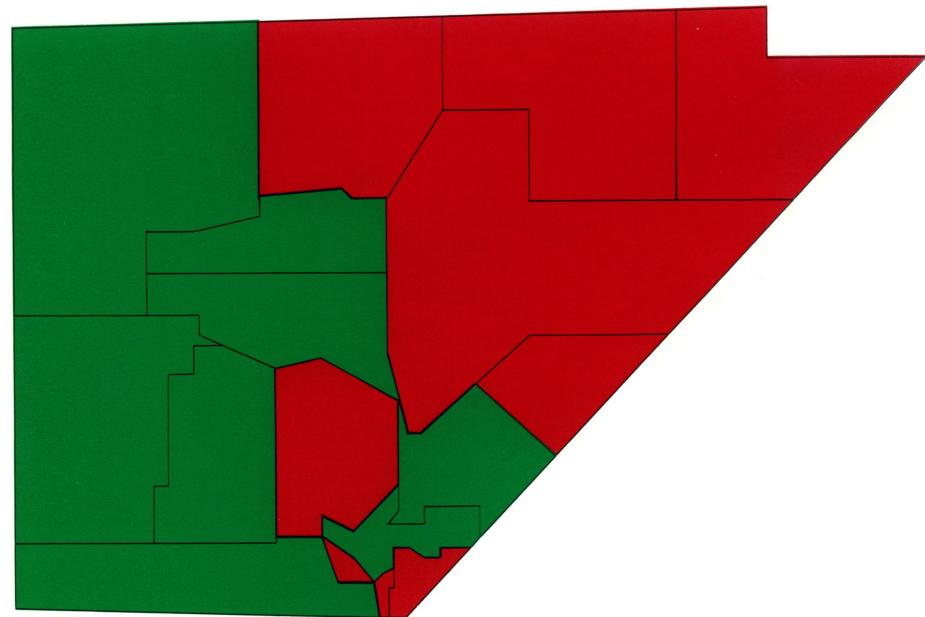
Color

Activity

Your Solution

Results from 1983 Experiment

- The following map of Nevada has been colored to indicate various geological features in each county.
- Estimate the larger land area-more red, more green, or the same-and mark your answer on the mentimeter poll in the next panel.
- **Please work fairly quickly, as if you were trying to gain an overall impression from a map.**



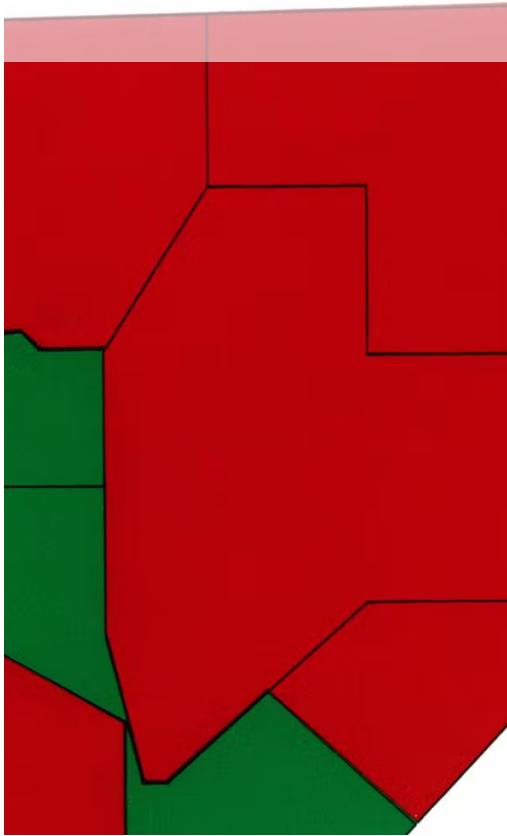
01 : 00

Color

Activity

Your Solution

Results from 1983 Experiment



Go to www.menti.com/g9wh7os8je

Which is the larger land-area?

 Mentimeter

01 : 00

Color

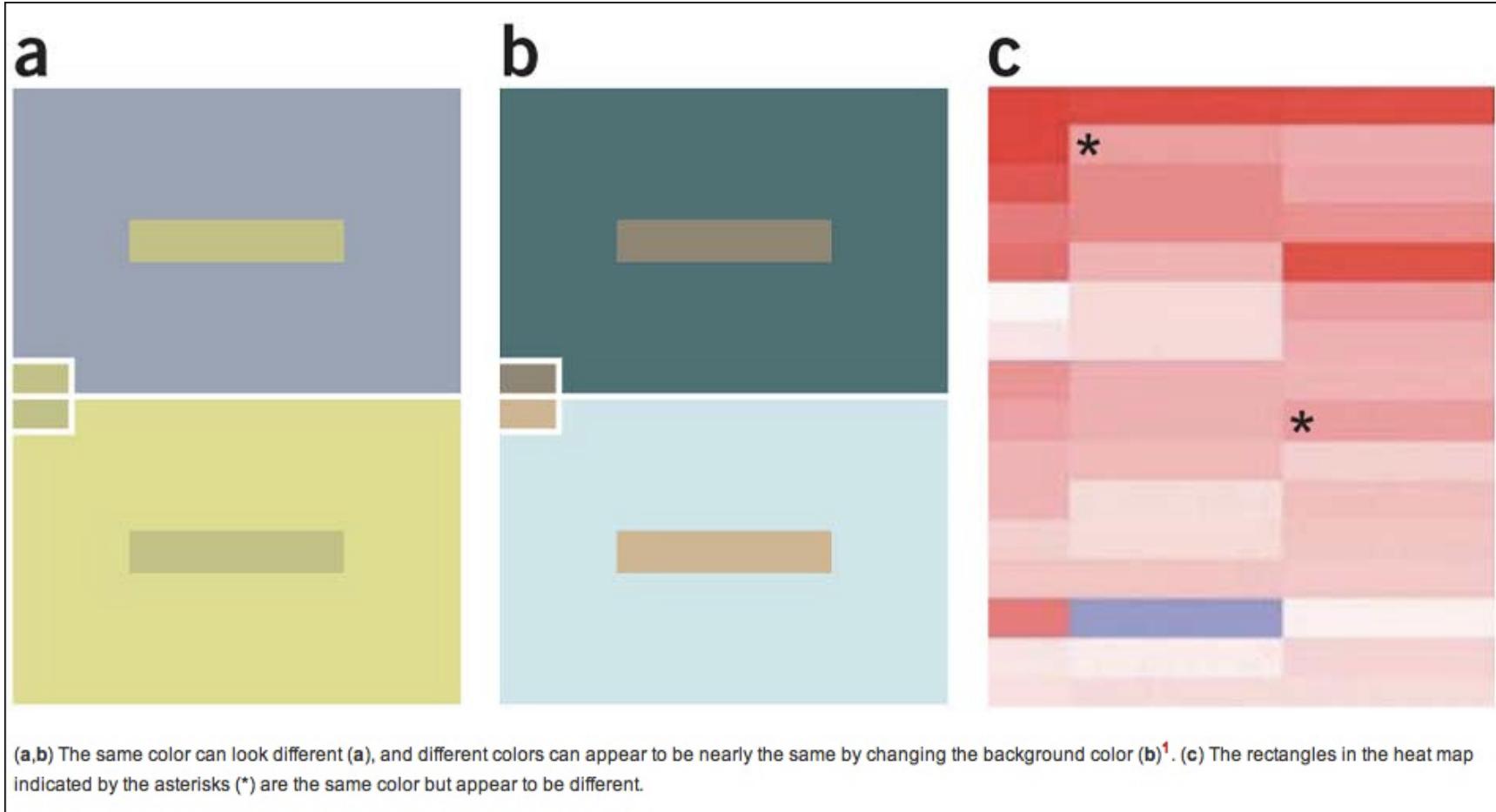
Activity

Your Solution

Results from 1983 Experiment

- The results from the original experiment were as follows

Simultaneous Contrast Affects Perception



Color Blindness

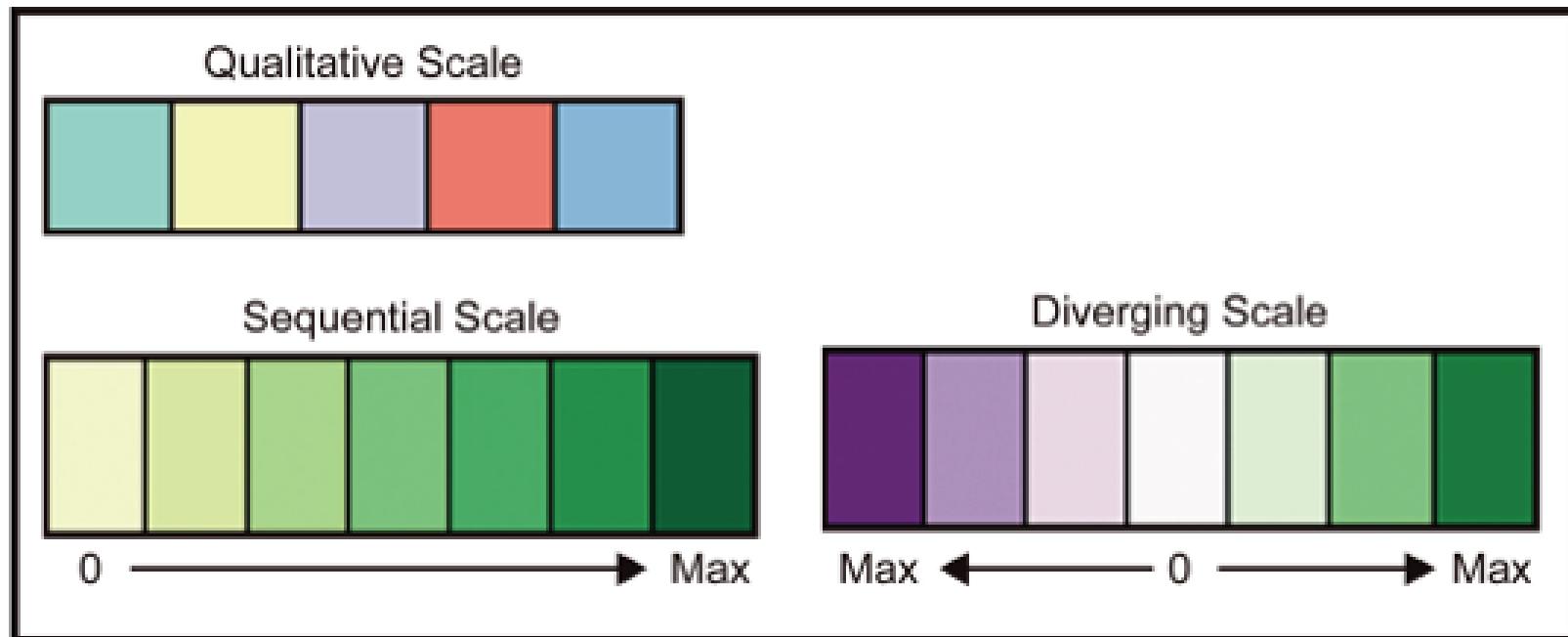
"About 1 in 12 men are color blind" -- NIH's At a glance: Color Blindness

Color Blindness Can Distort a Person's Reading/Interpretation of a Chart

- **Personal Check:** To have a feel for color blindness (if you are not color blind), you can take this color blind test
- Given the high prevalence of color blind individuals, your charts **should** accommodate for color-blindness. **How?**
 - Use color sparingly
 - Use color friendly palettes, e.g., see <https://colorbrewer2.org/>

Color Brewer: Color Scales and their Selection

Nominal



Source: Brewer, Cynthia A. "Color use guidelines for data representation." Proceedings of the Section on Statistical Graphics, American Statistical Association. 1999.

Recap

Summary of Main Points

By now, you should be able to do the following:

- Explain the concept of "graphical excellence"
- Explain the theory of data graphics
- Optimize visual encoding based on data types
- Understand why color should be used sparingly and how to select appropriate colors (when color is a must)