ISA 401/501: Business Intelligence & Data Visualization

13: Fundamentals of Data Visualization

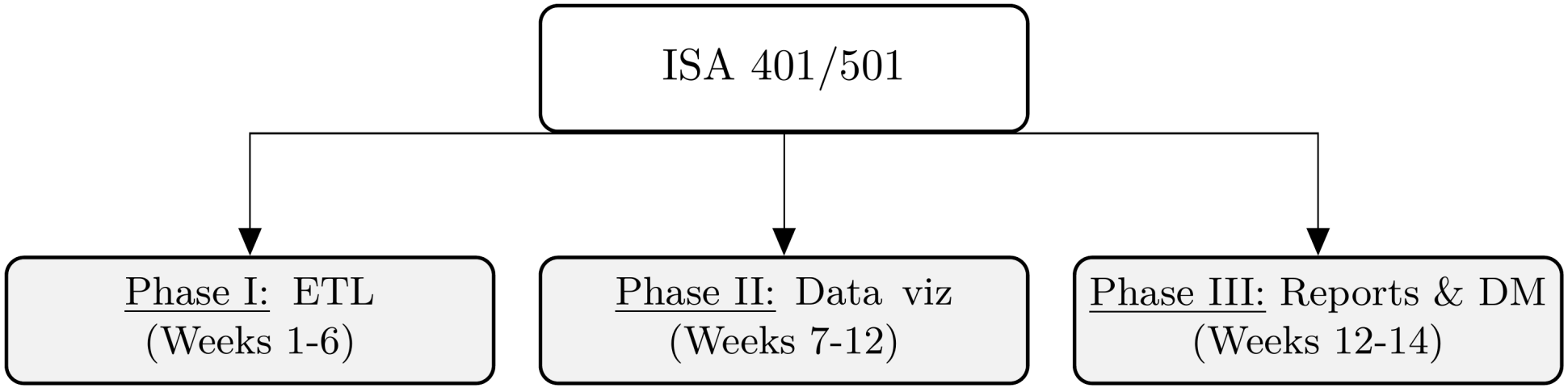
Fadel M. Megahed, PhD

Enders Associate Professor
Farmer School of Business
Miami University

@FadelMegahed
fmegahed
fmegahed@miamioh.edu
Automated Scheduler for Office Hours

Fall 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

Non-graded activity: Terrible Charts

05:00

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.

Non-graded activity: Terrible Charts

05:00

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.

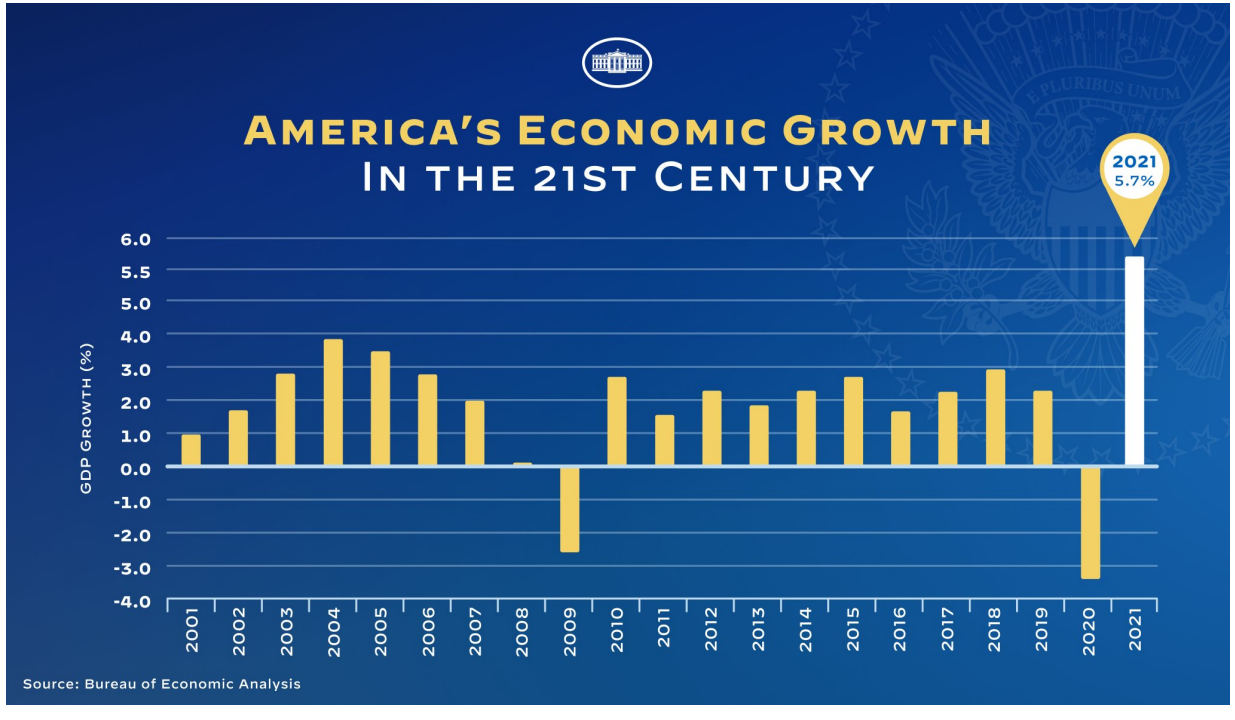
Main Issue(s): (Insert below)

Non-graded activity: Terrible Charts

05:00

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 corrected.



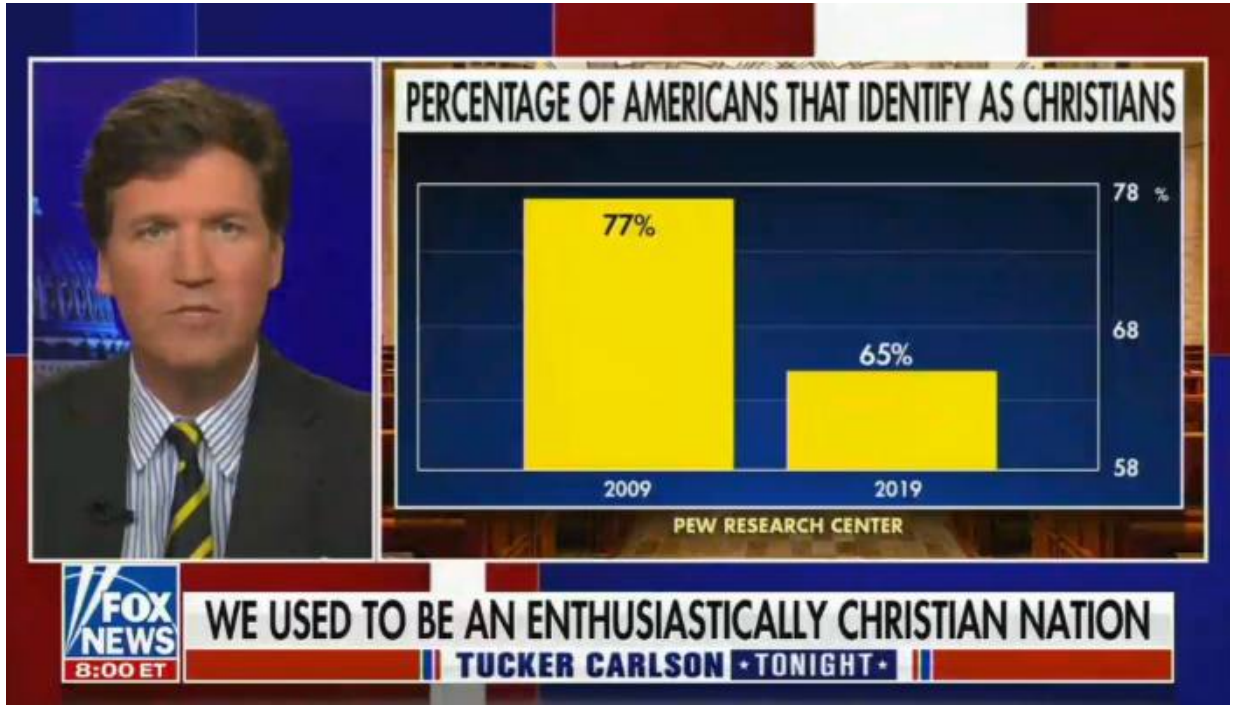
Main Issue(s): (Insert below)

Non-graded activity: Terrible Charts

05:00

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**

Source: Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, Conn: Graphics Press, P. 13.

Show/Reveal the Data: Anscombe's Dataset

In a seminal paper, Anscombe stated:

> **Few of us escape being indoctrinated with these notions:**
>
> - numerical **calculations are exact, but graphs are rough**;
> - for any particular kind of **statistical data there is just one set of calculations constituting a correct statistical analysis**;
> - performing **intricate calculations is virtuous**, whereas **actually looking at the data is cheating**.

He proceeded by stating that

> a computer should **make both calculations and graphs**. Both sorts of output should be studied; each will contribute to understanding.

Now, let us consider his four datasets, each consisting of eleven (x,y) pairs.

Show/Reveal the Data: Anscombe's Dataset

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.1	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.1	5.39	12.5
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

Showing 1 to 11 of 11 entries

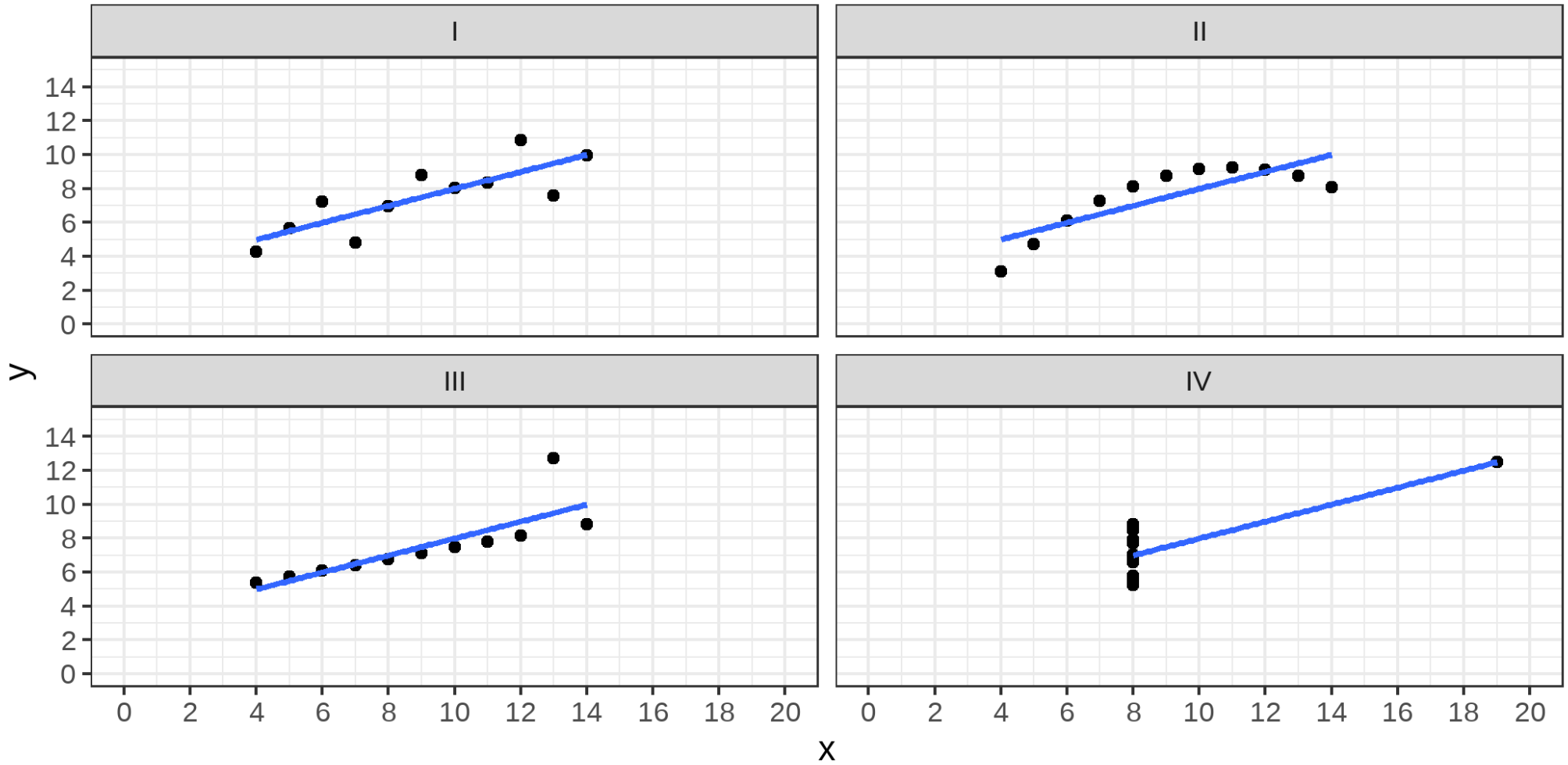
Show/Reveal the Data: Anscombe's Dataset

set	x.mean	x.sd	y.mean	y.sd	corr
I	9	3.32	7.5	2.03	0.82
II	9	3.32	7.5	2.03	0.82
III	9	3.32	7.5	2.03	0.82
IV	9	3.32	7.5	2.03	0.82

Showing 1 to 4 of 4 entries

Previous 1 Next

Show/Reveal the Data: Anscombe's Dataset

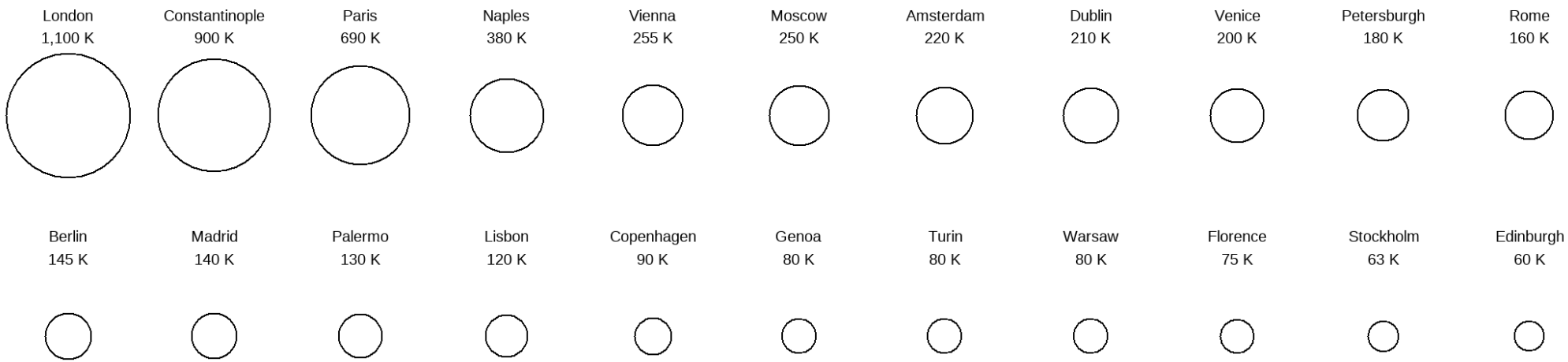


Substance

05:00

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.



Source: Chart made by Fadel M. Megahed based on data from http://www.stat.uiowa.edu/~luke/data/Playfair

Substance

05:00

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.

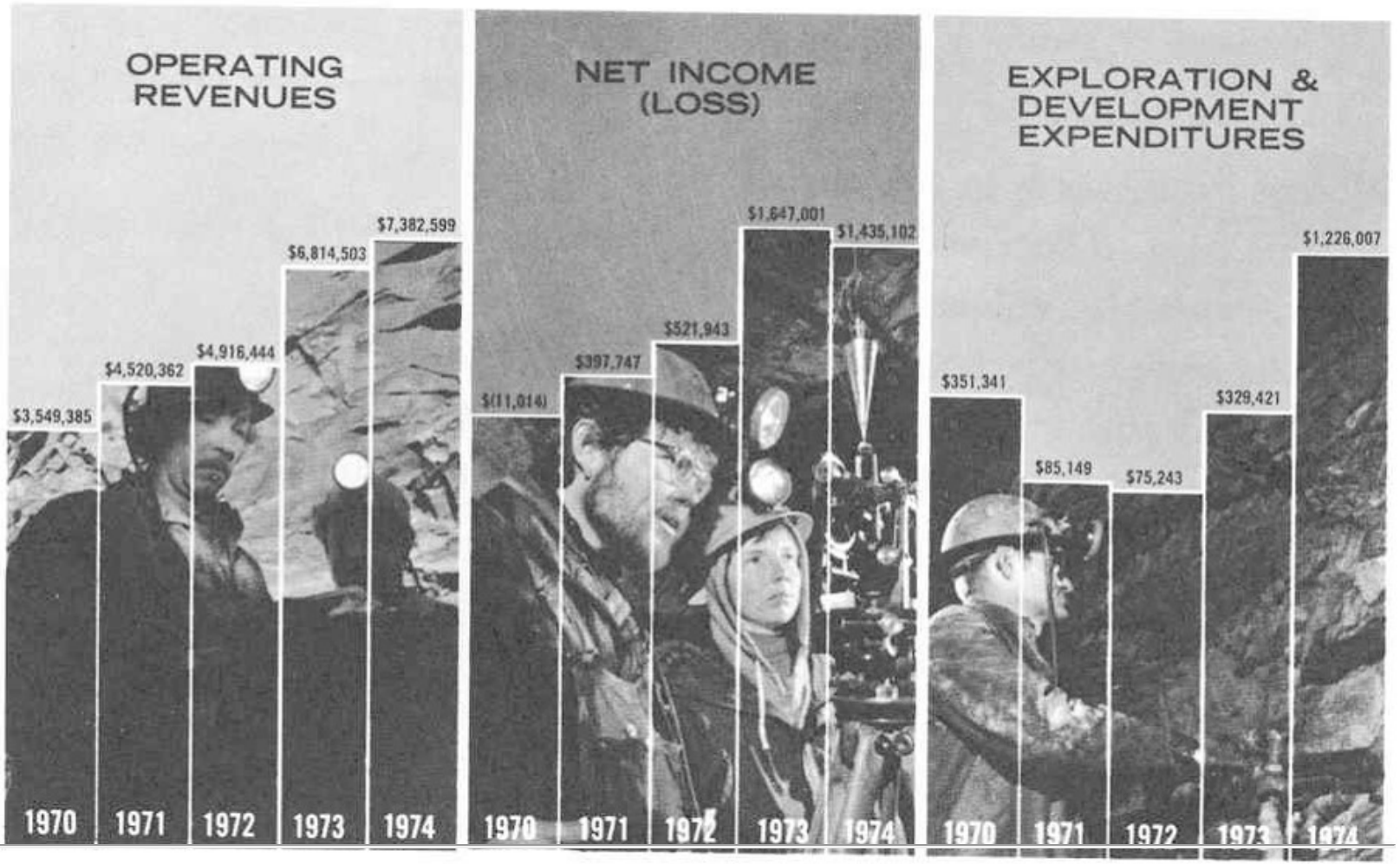
Substance

05:00

Activity | Your Solution | My Solution

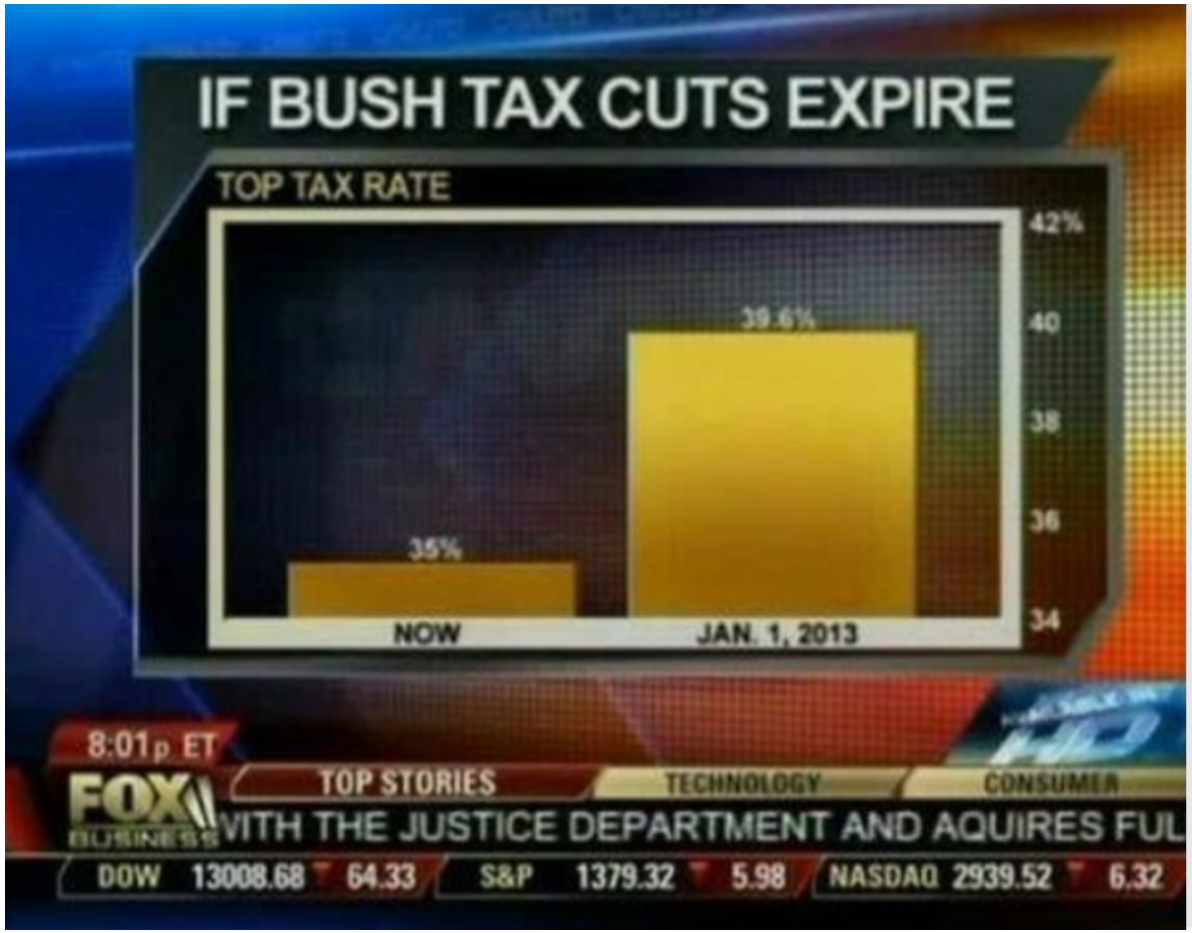
- In my opinion, a **dot chart is more effective** than the bubble chart. The **population would be mainly encoded using the position**; you can still use area as a secondary encoding mechanism.

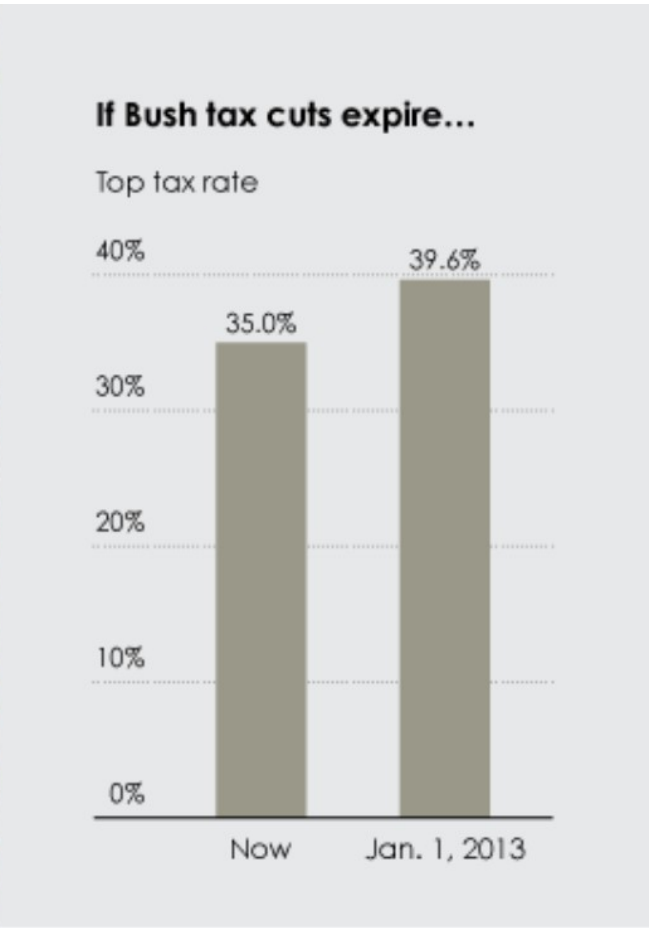
Avoid Distortion of Data



Source: Graph from Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, Conn: Graphics Press, P. 54.

Avoid Distortion of Data

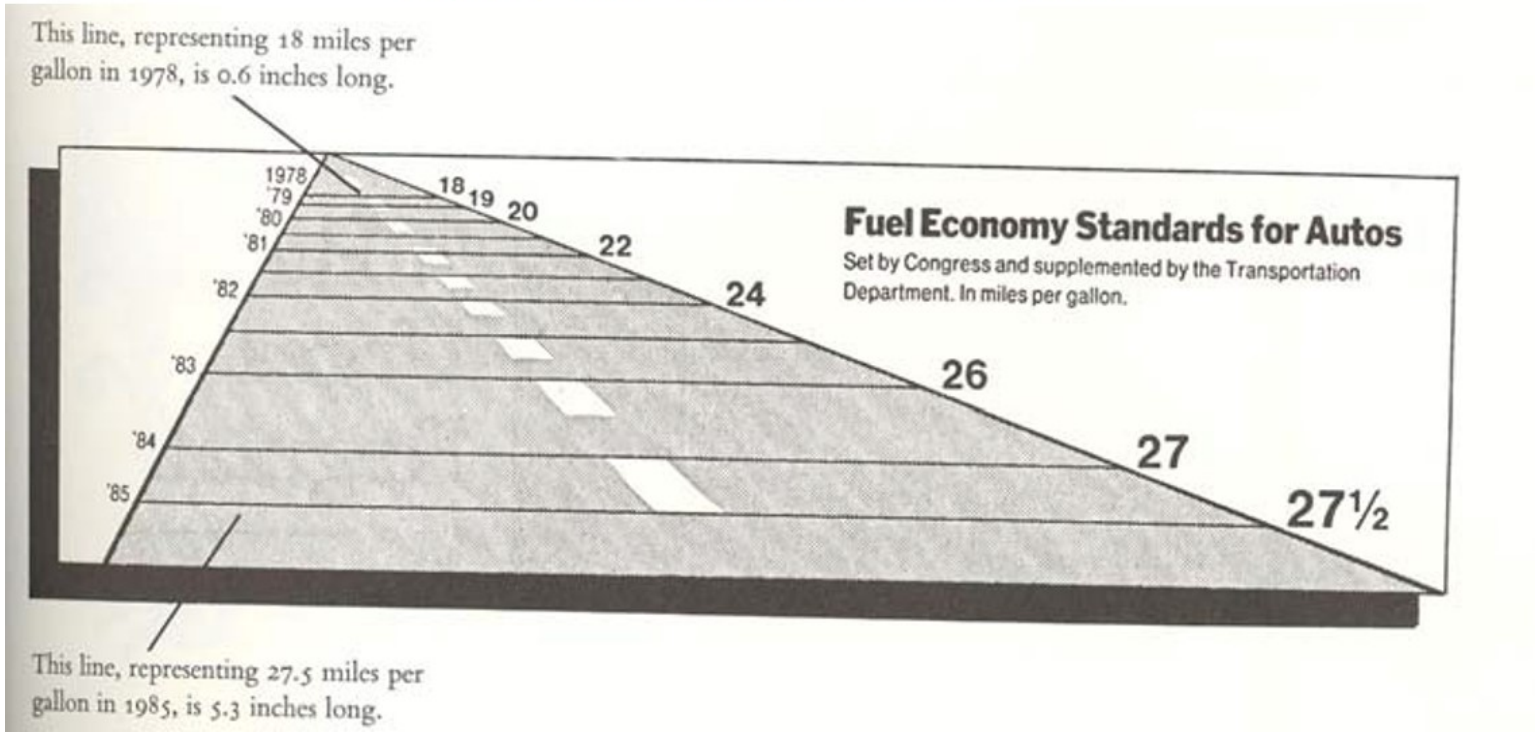




Source: Nathan Yau (2012). Fox News continues charting excellence.

Avoid Distortion of Data: The Lie Factor

$$\frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$



Source: Graph from Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, Conn: Graphics Press, P. 57.

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.

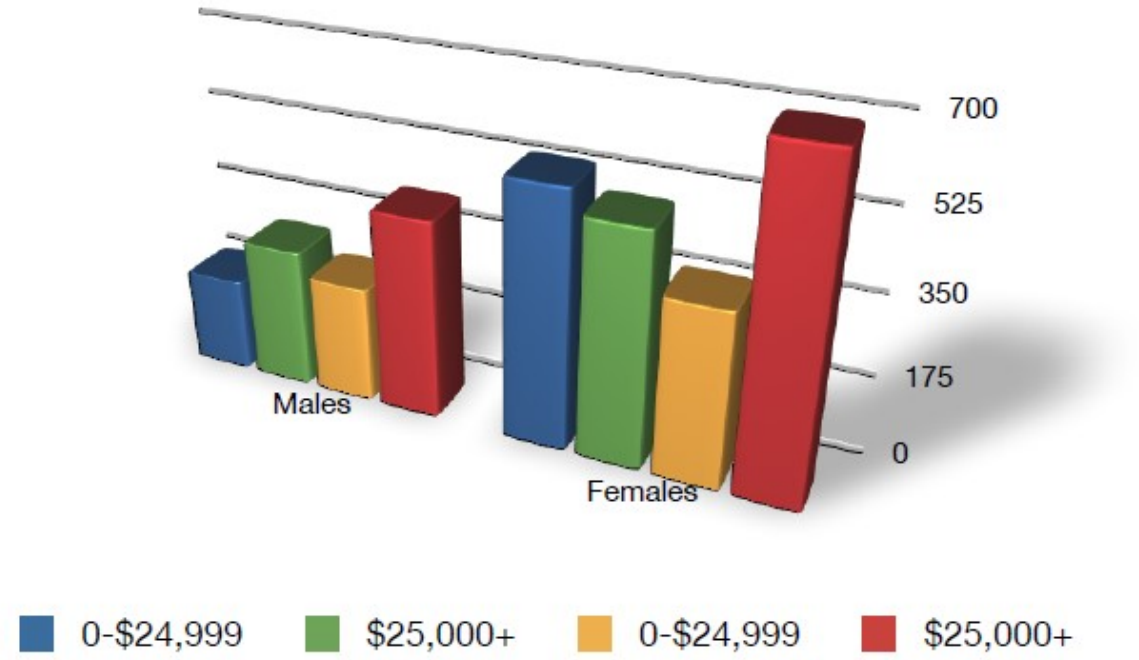
$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

$$= \text{proportion of a graphic's ink devoted to the non-redudant display of data-information}$$

$$= 1.0 - \text{proportion of a graphic that can be erased}$$

Definition of Data Ink

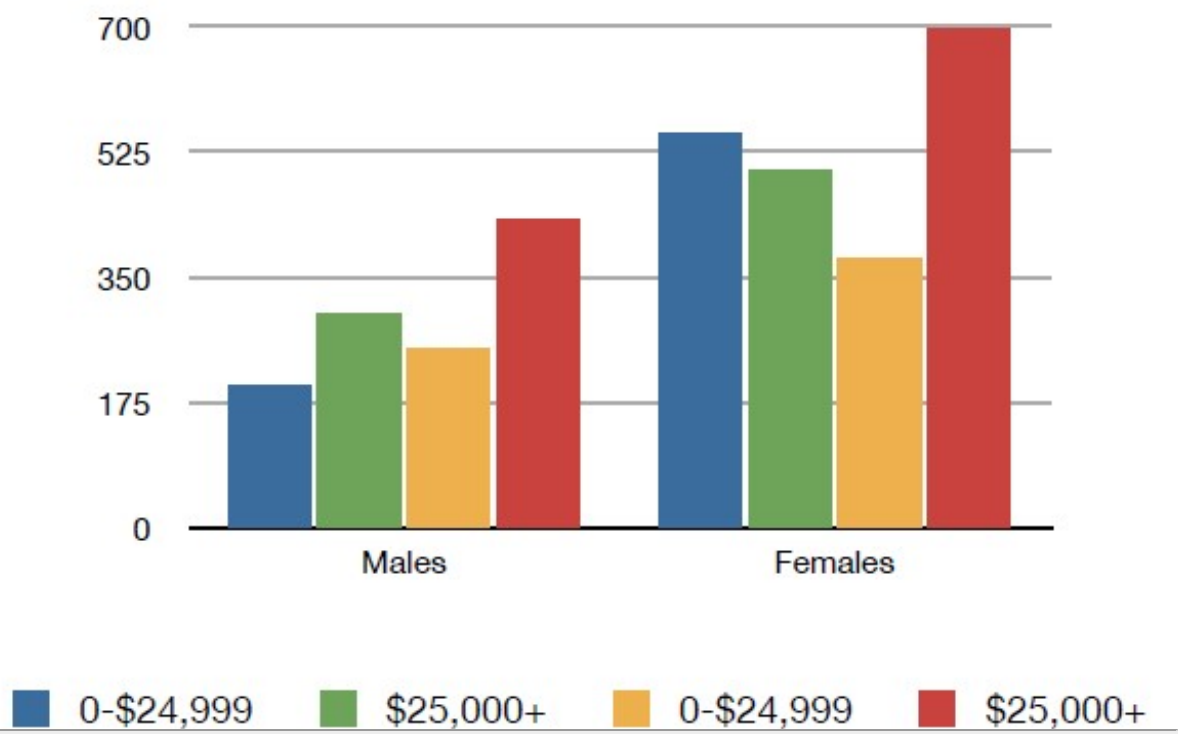
- **Data-ink refers to the non-erasable ink used for presenting the data**.
  - If data-ink would be removed from the image, the graphic would lose the content.



Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/

Definition of Data Ink

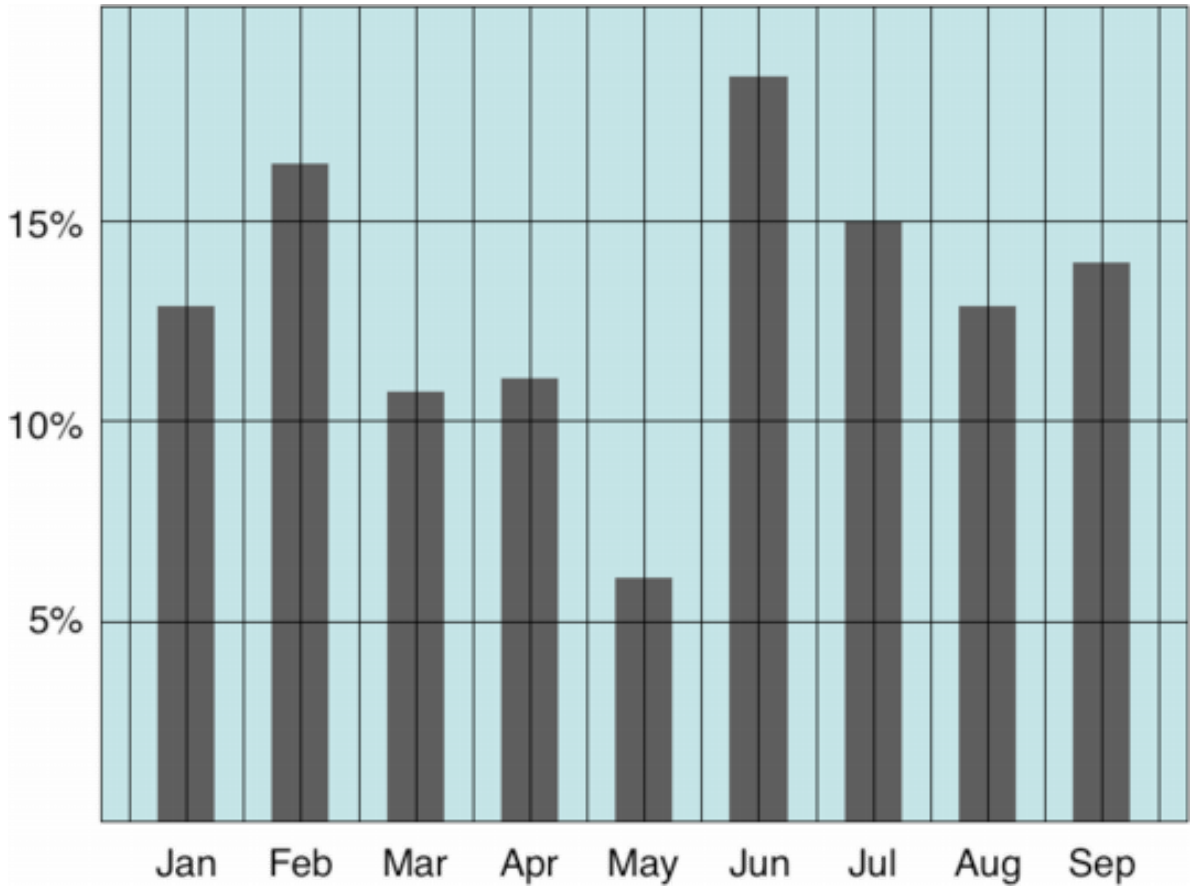
- **Data-ink refers to the non-erasable ink used for presenting the data**.
  - If data-ink would be removed from the image, the graphic would lose the content.



Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/

Focus on Data: Avoid Chartjunk

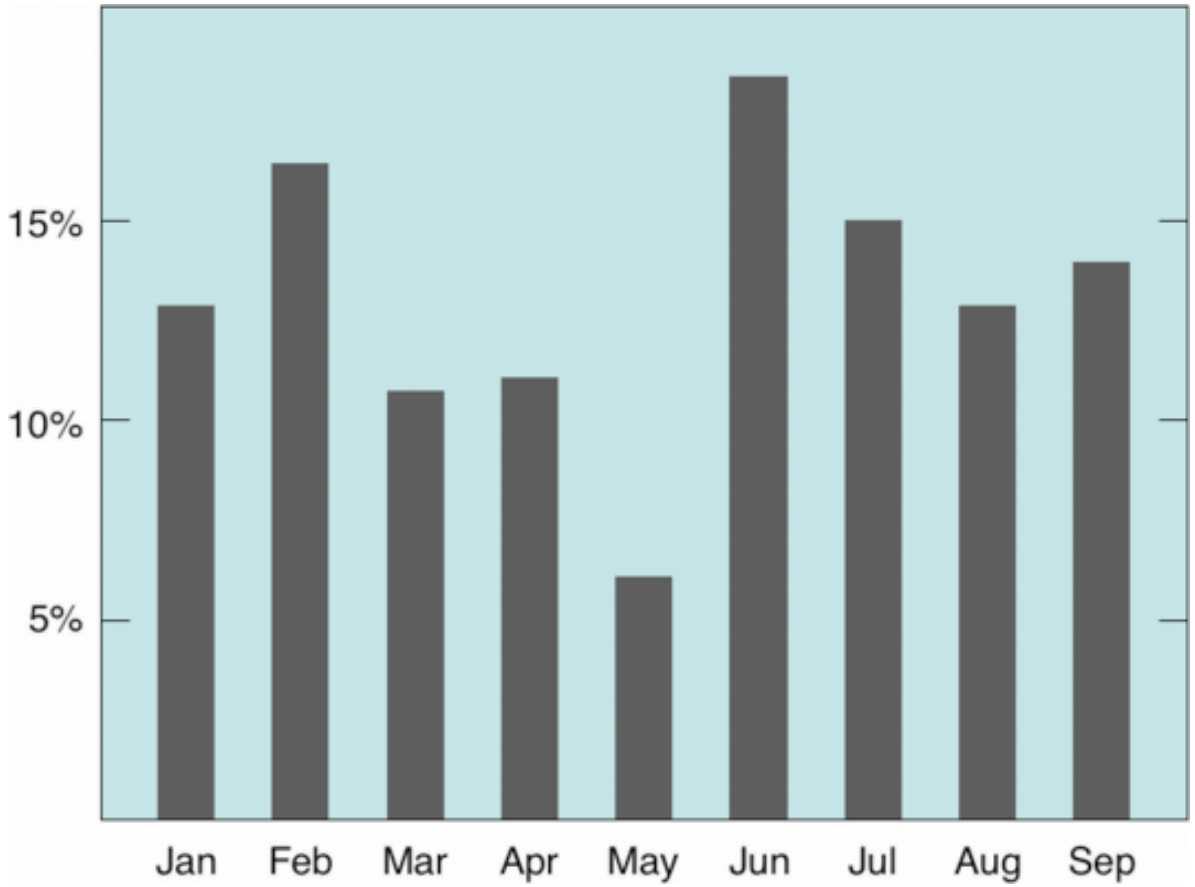
Chartjunk is the extraneous visual elements that distract from the message!!



Source: The image is from The Data-Ink Ratio Example

Focus on Data: Avoid Chartjunk

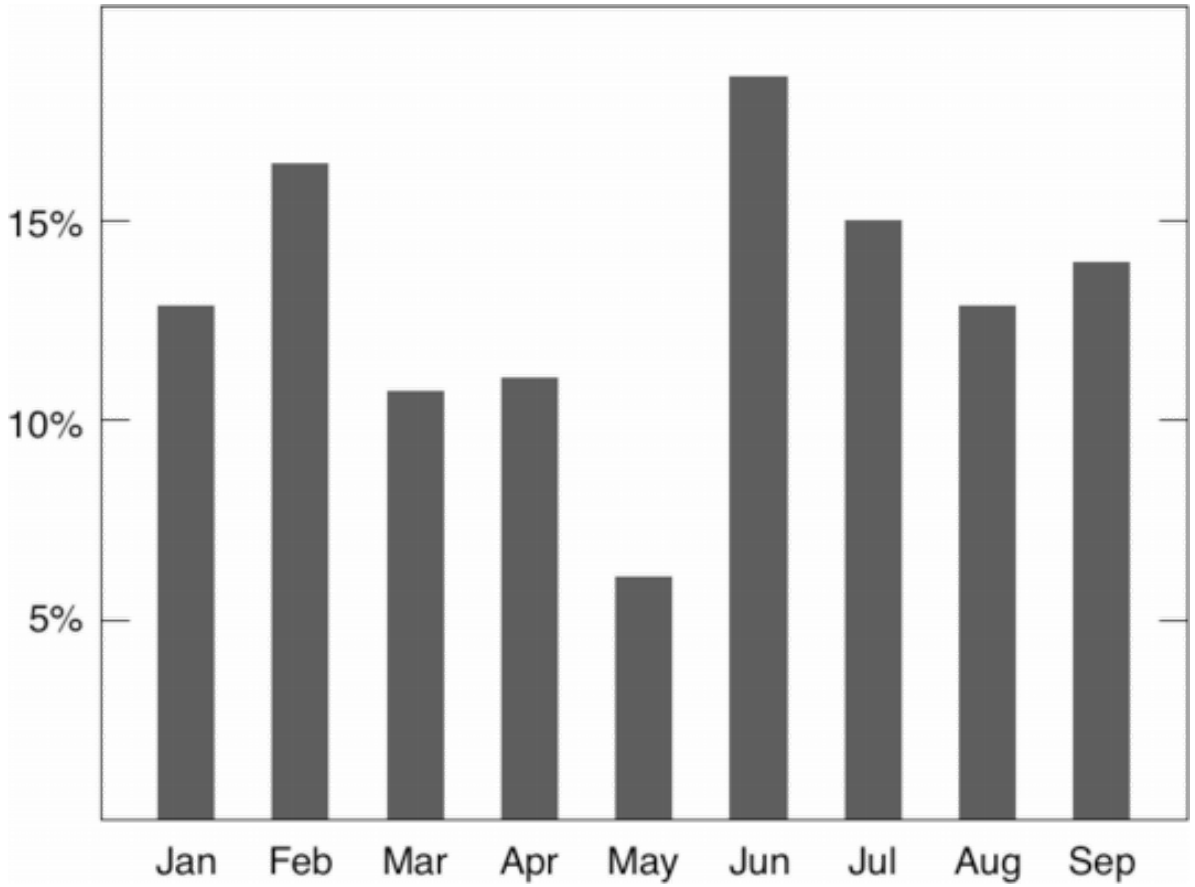
Chartjunk is the extraneous visual elements that distract from the message!!



Source: The image is from The Data-Ink Ratio Example

Focus on Data: Avoid Chartjunk

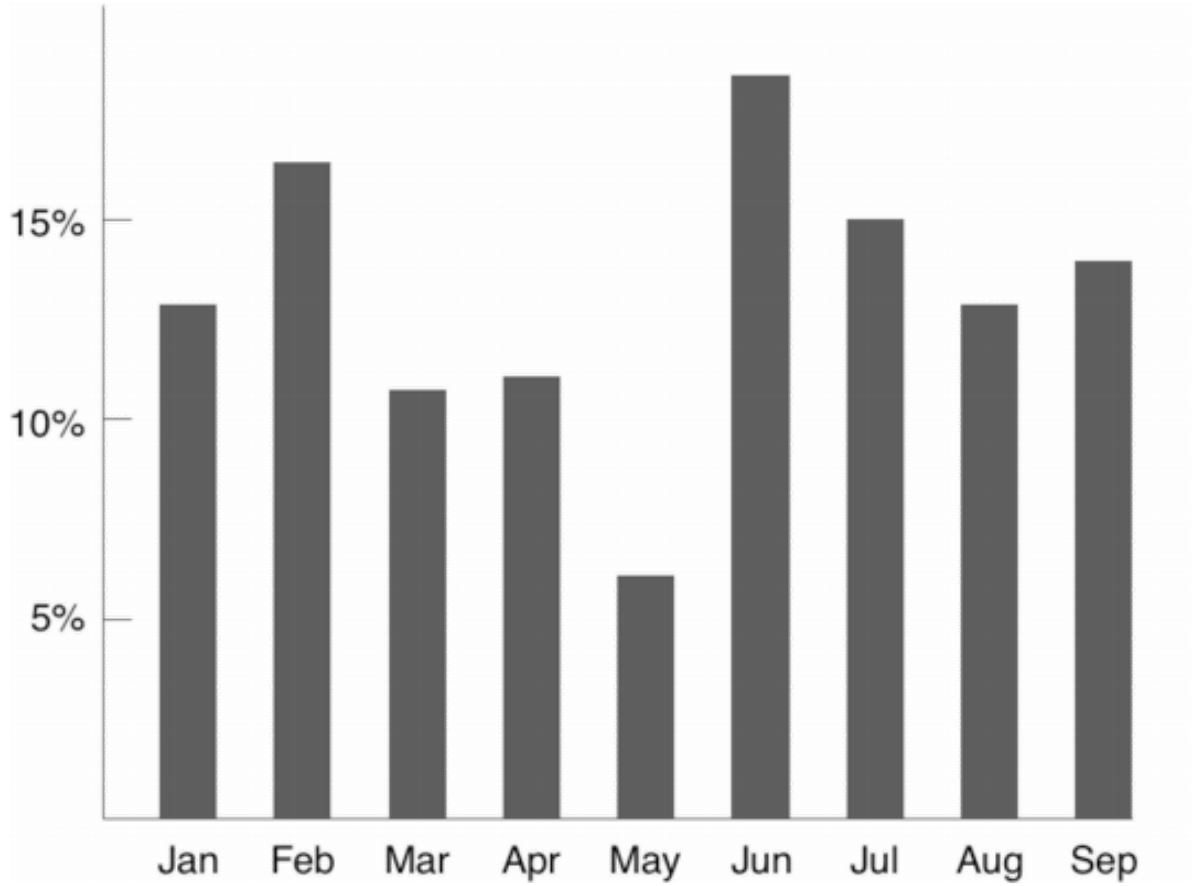
Chartjunk is the extraneous visual elements that distract from the message!!



Source: The image is from The Data-Ink Ratio Example

Focus on Data: Avoid Chartjunk

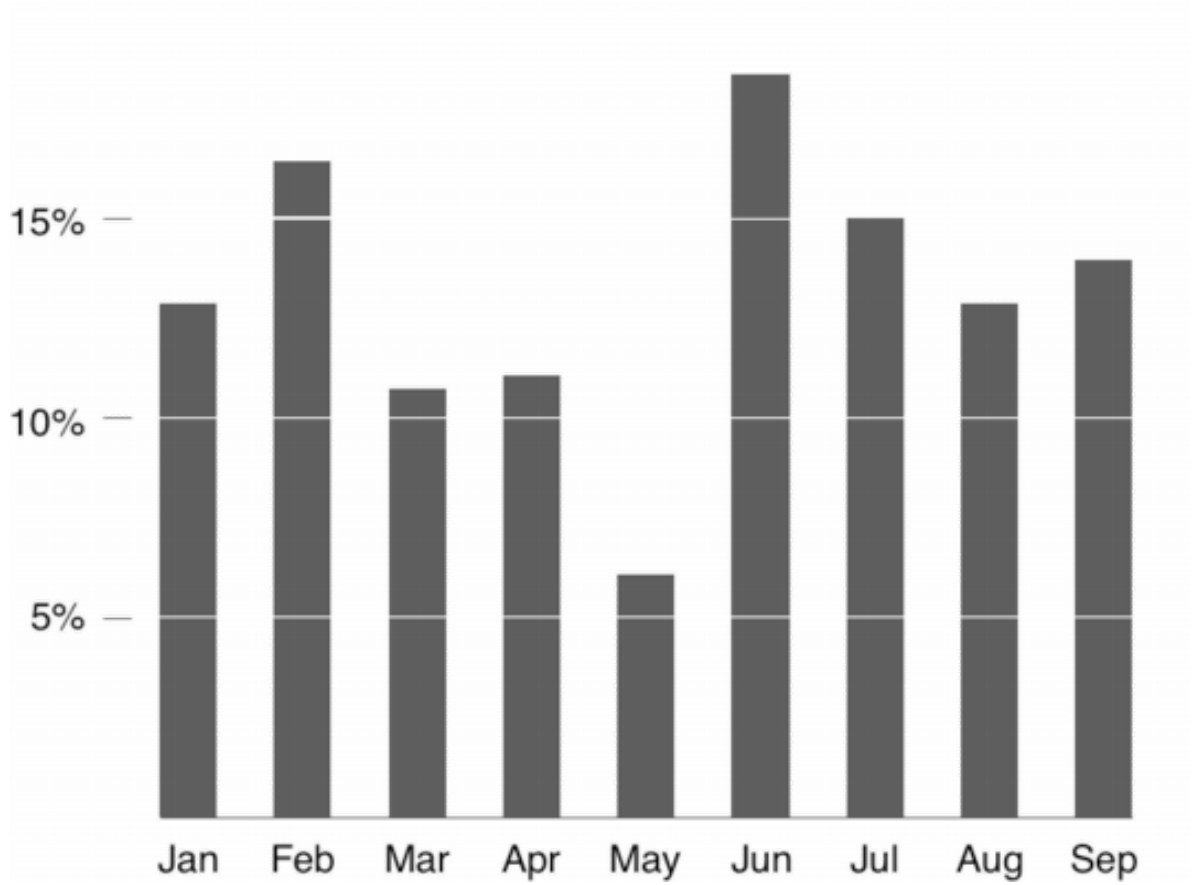
Chartjunk is the extraneous visual elements that distract from the message!!



Source: The image is from The Data-Ink Ratio Example

Focus on Data: Avoid Chartjunk

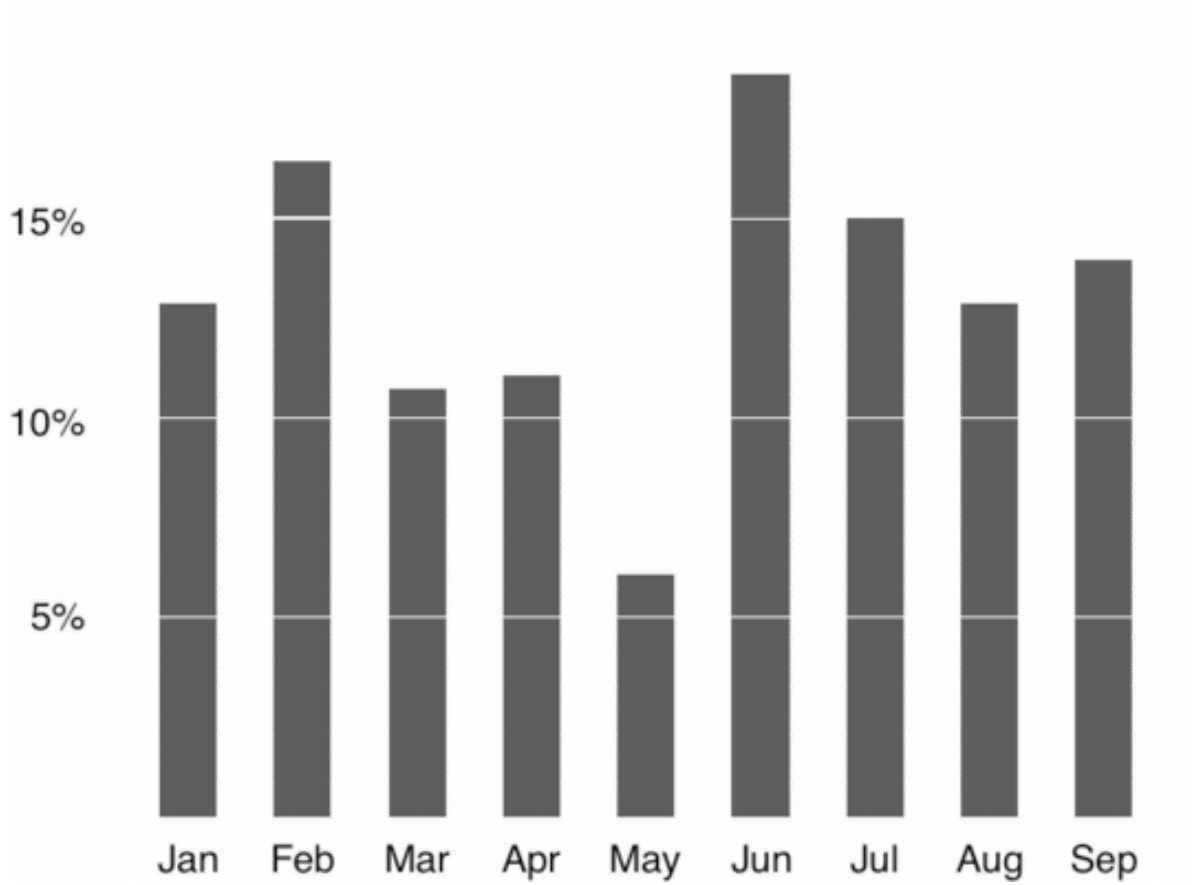
Chartjunk is the extraneous visual elements that distract from the message!!



Source: The image is from The Data-Ink Ratio Example

Focus on Data: Avoid Chartjunk

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



Source: The image is from The Data-Ink Ratio Example

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

Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/. Information is based on comments by Pat Hanrahan in November 2007. Those can be found at http://www.perceptualedge.com/blog/?p=199

Data Models

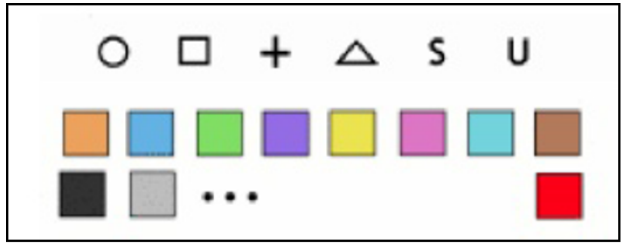
Data Types (from S. Stevens, Theory of Scales)

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

Data Types: Explained

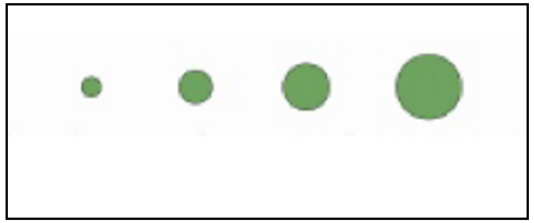
Nominal:

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



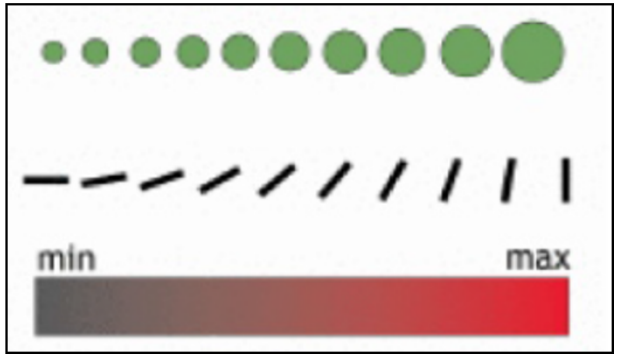
Ordinal:

- 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, ...
- Origin is meaningful, we can compute ratios, proportions, differences, etc.

Non-Graded Activity: Data Terminology

02:00

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

Non-Graded Activity: Data Terminology

02:00

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

Marks Points Lines Areas

Channels

Position

Size

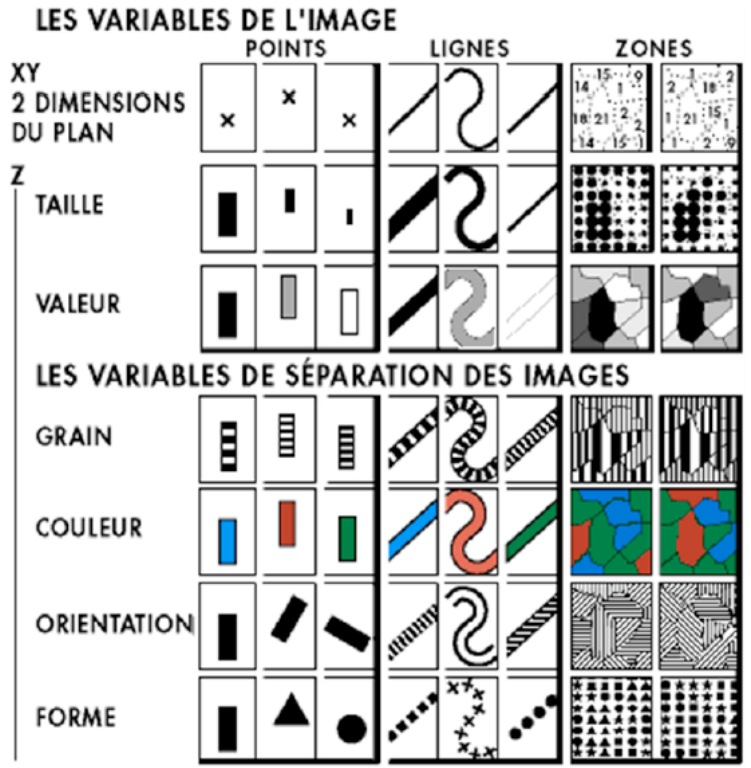
(Grey)Value

Texture

Color

Orientation

Shape



Source: Bertin (1967), Seminology of Graphics.

Mapping to Data Types

	Nominal	Ordinal	Quantitative
Position	✔	✔	✔
Size	✔	✔	~
(Grey)Value	✔	✔	~
Texture	✔	~	✖
Color	✔	✖	✖
Orientation	✔	✖	✖
Shape	✔	✖	✖

✔ = Good
~ = OK
✖ = Bad

Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/

Visual Channels and their Precision

Source: Franconeri, et al. (2021). The Science of Visual Data Communication: What Works. To view the full size image, click here.

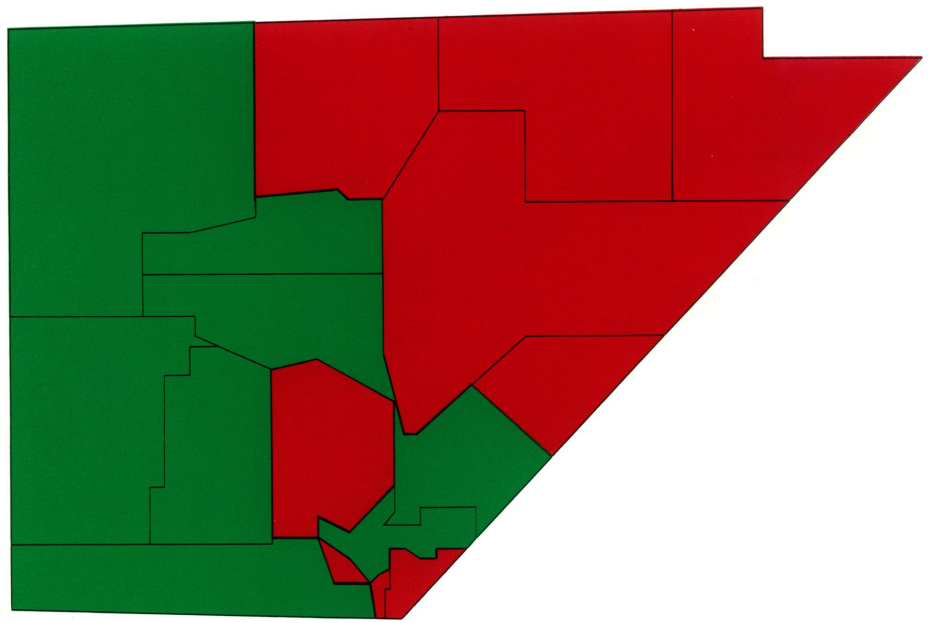
Color Should Be Used Sparingly

Color

01:00

Activity | Your Solution | Results from 1983 Experiment

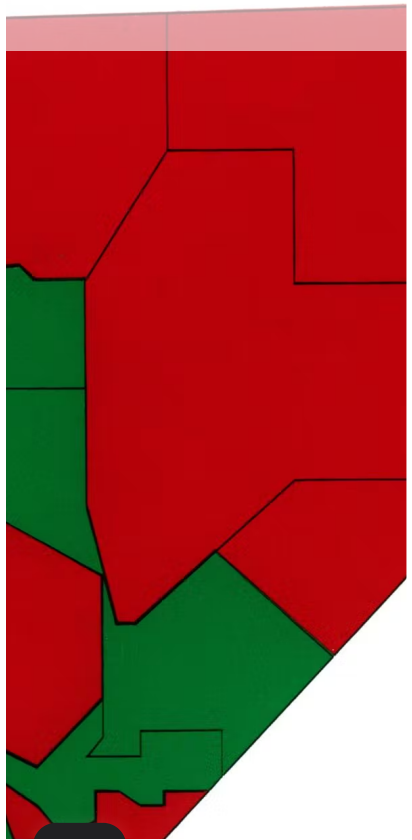
- The following map of Nevada has been colored to indicate various geological features in each county.
- Estimate the larger land area-more red, more green, or the same-and mark your answer on the mentimeter poll in the next panel.
- **Please work fairly quickly, as if you were trying to gain an overall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



Color

01:00

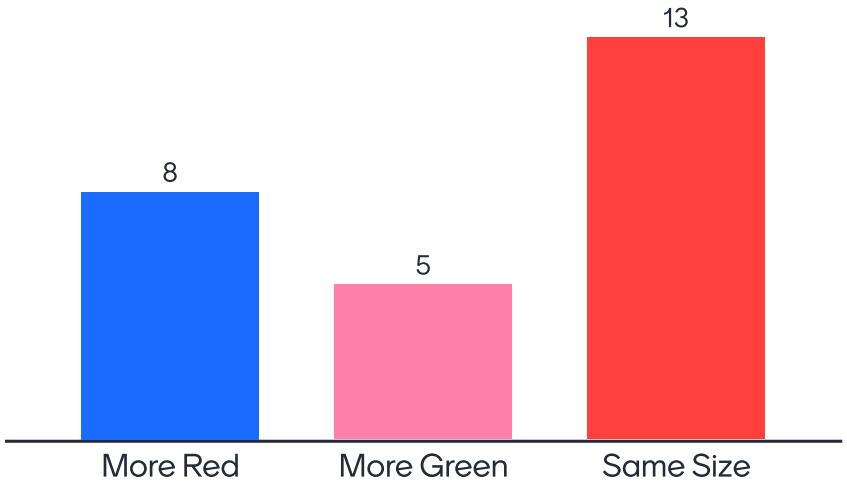
Activity | Your Solution | Results from 1983 Experiment



Go to **www.menti.com/g9wh7os8je**

Which is the larger land-area?

Mentimeter



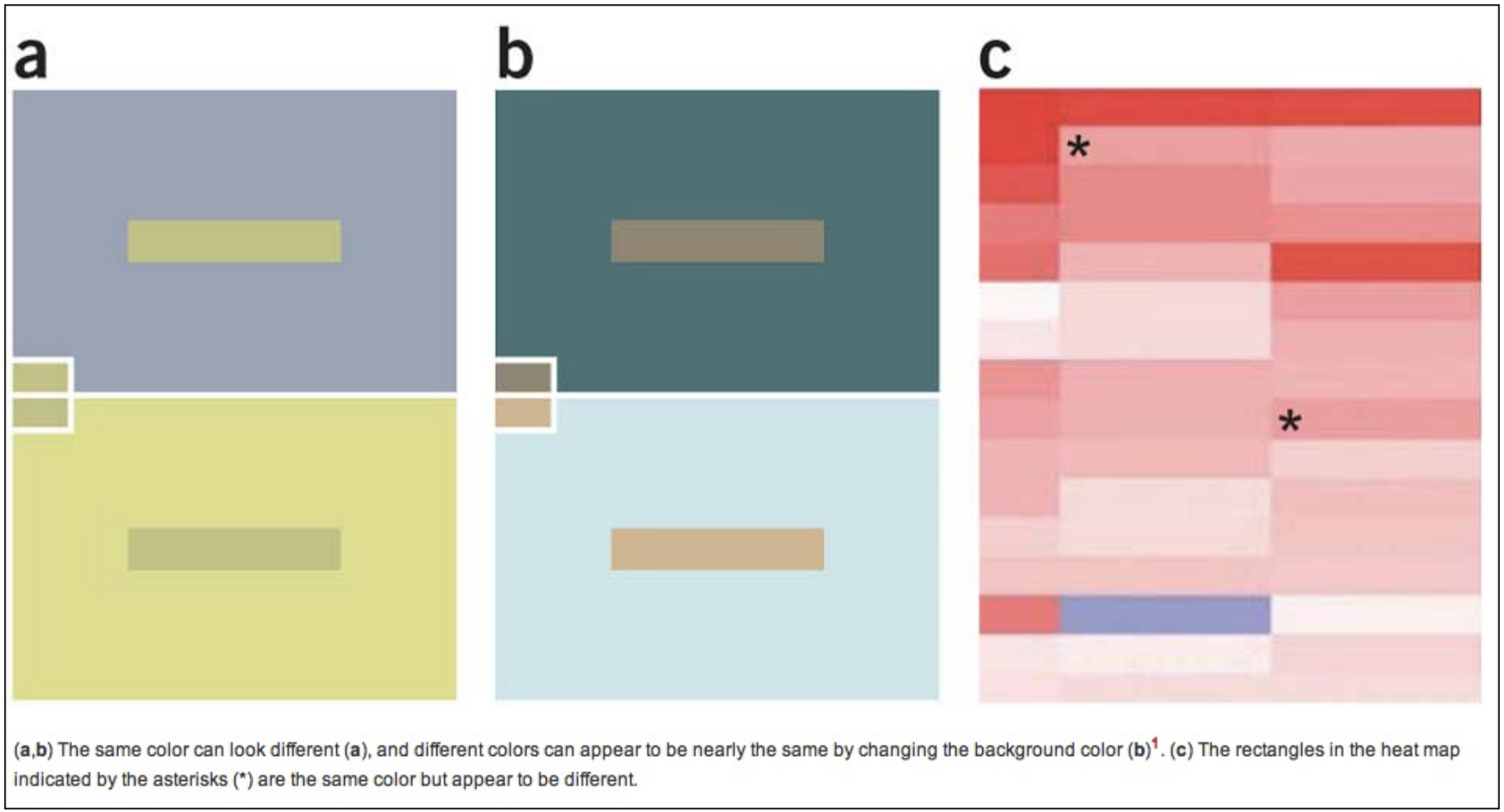
Color

01:00

Activity | Your Solution | Results from 1983 Experiment

- The results from the original experiment were as follows

Simultaneous Contrast Affects Perception



(**a**,**b**) The same color can look different (**a**), and different colors can appear to be nearly the same by changing the background color (**b**)[1]. (**c**) The rectangles in the heat map indicated by the asterisks (*) are the same color but appear to be different.

Source: Slide Adapted from Hanspeter Pfister. Lecture Notes for CS 171. Harvard University. http://cs171.org/

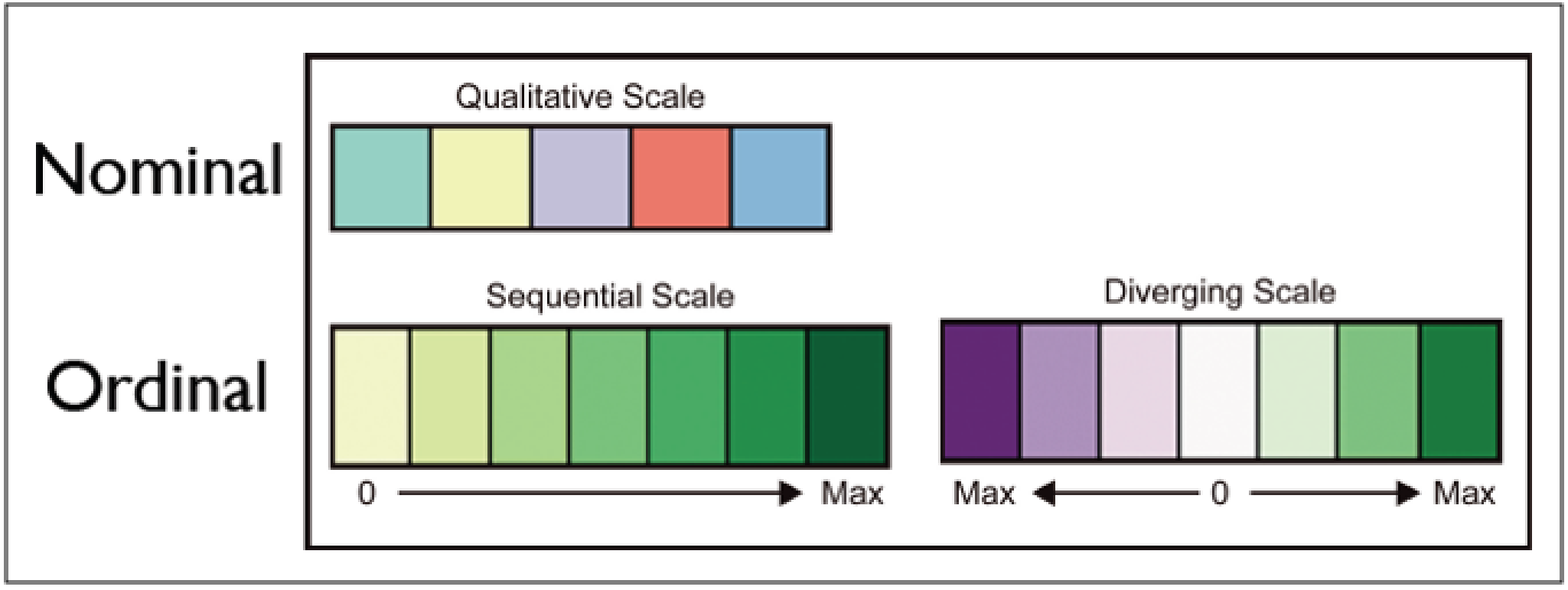
Color Blindness

> "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



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)

Things to Do Prior to Next Class

Please go through the following two supplementary readings and complete assignment 11.

- The Lie Factor and the Baseline Paradox; especially noting what the authors mean by "baseline", how the lie factor may be ignored in time-series applications, and/or in applications involving a "ratio" scale.
- Useful junk? The effects of visual embellishment on comprehension and memorability of charts, which presents an experimental counter against Tufte's argument for simplicity (by quantifying vividness and recall of data from the more artistic charts). Note they define **"ratio"** different from how we have defined in class.