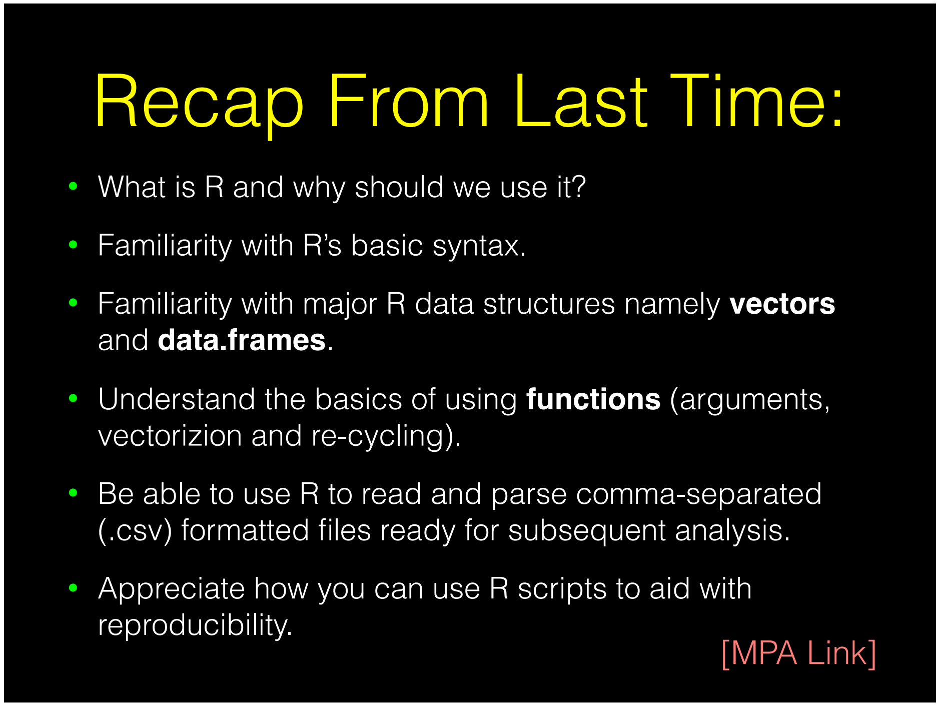
BGGN 213

Data visualization with R

Barry Grant

UC San Diego

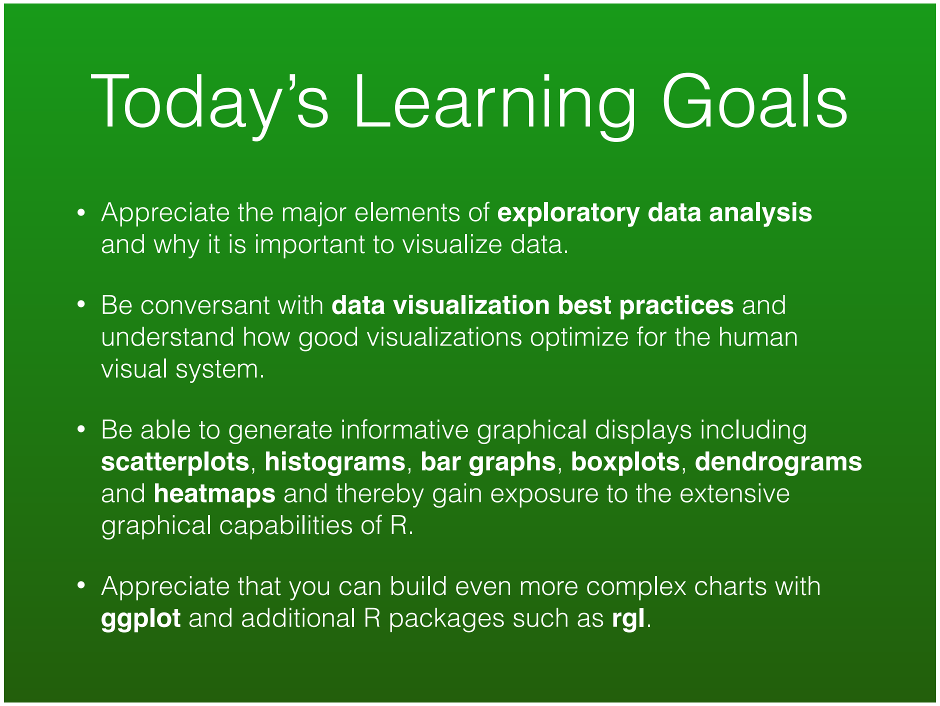
http://thegrantlab.org/bggn213



Recap From Last Time:

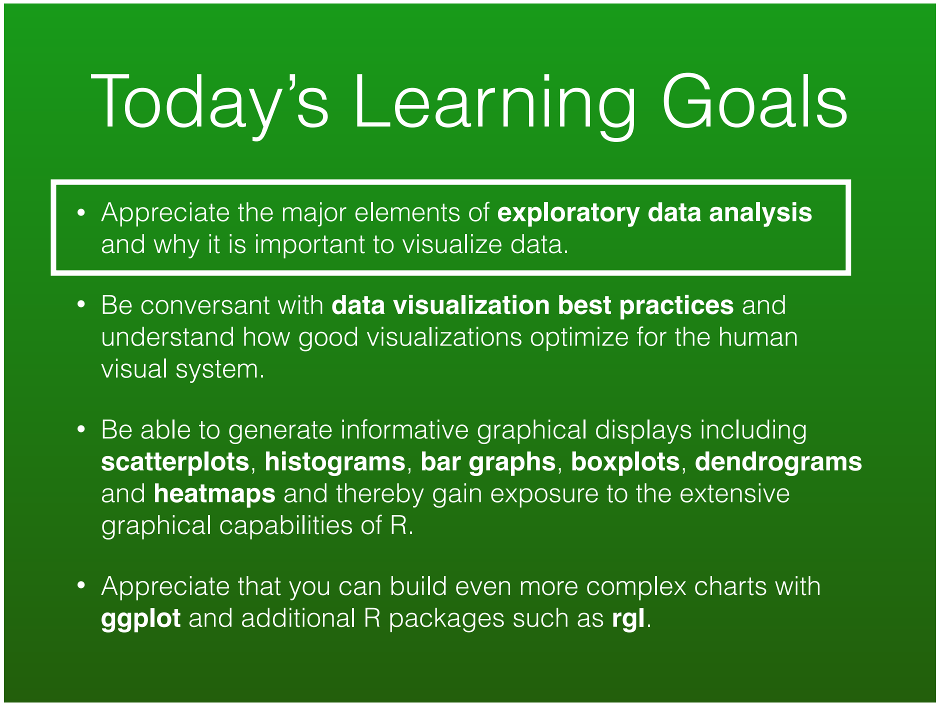
- What is R and why should we use it?
- Familiarity with R’s basic syntax.
- Familiarity with major R data structures namely **vectors** and **data.frames**.
- Understand the basics of using **functions** (arguments, vectorizion and re-cycling).
- Be able to use R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.
- Appreciate how you can use R scripts to aid with reproducibility.

[MPA Link]



Today’s Learning Goals

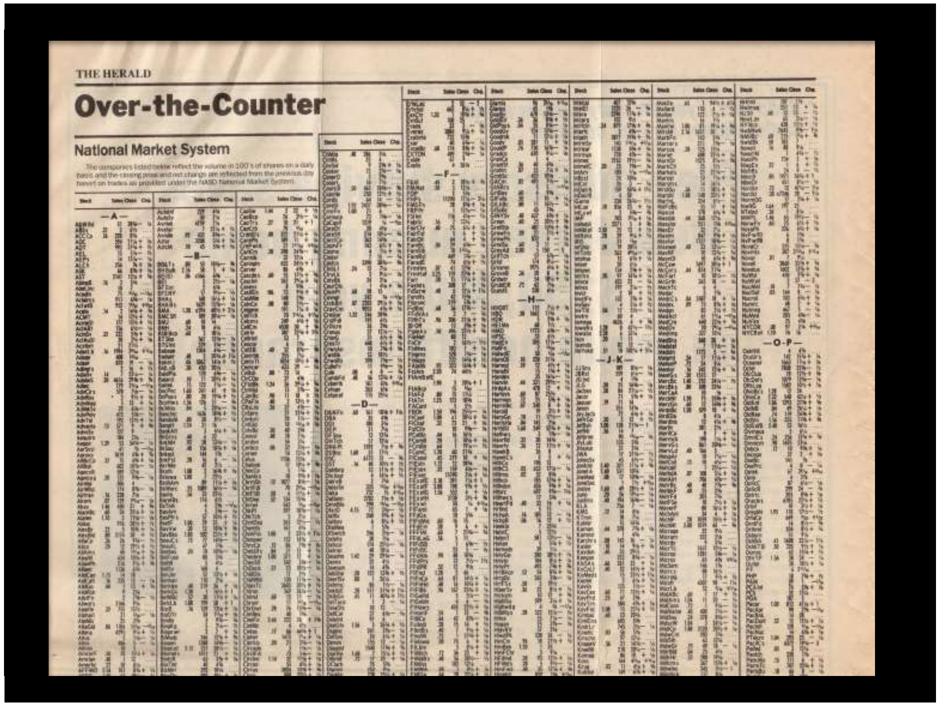
- Appreciate the major elements of **exploratory data analysis** and why it is important to visualize data.
- Be conversant with **data visualization best practices** and understand how good visualizations optimize for the human visual system.
- Be able to generate informative graphical displays including **scatterplots**, **histograms**, **bar graphs**, **boxplots**, **dendrograms** and **heatmaps** and thereby gain exposure to the extensive graphical capabilities of R.
- Appreciate that you can build even more complex charts with **ggplot** and additional R packages such as **rgl**.



Today’s Learning Goals

- Appreciate the major elements of **exploratory data analysis** and why it is important to visualize data.
- Be conversant with **data visualization best practices** and understand how good visualizations optimize for the human visual system.
- Be able to generate informative graphical displays including **scatterplots**, **histograms**, **bar graphs**, **boxplots**, **dendrograms** and **heatmaps** and thereby gain exposure to the extensive graphical capabilities of R.
- Appreciate that you can build even more complex charts with **ggplot** and additional R packages such as **rgl**.

Why visualize at all?



THE HERALD

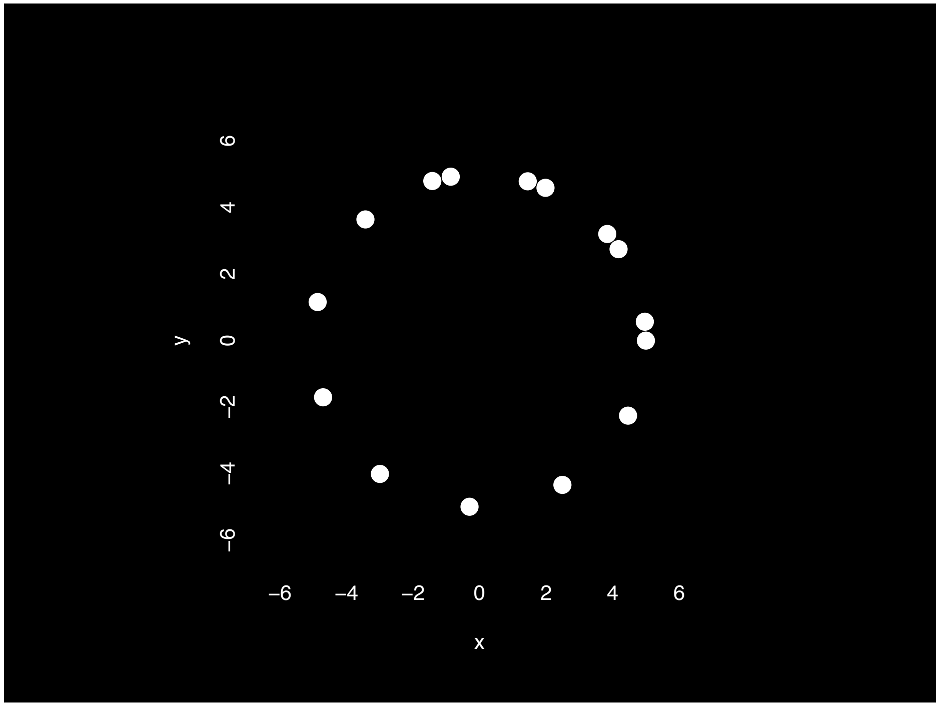
Over-the-Counter

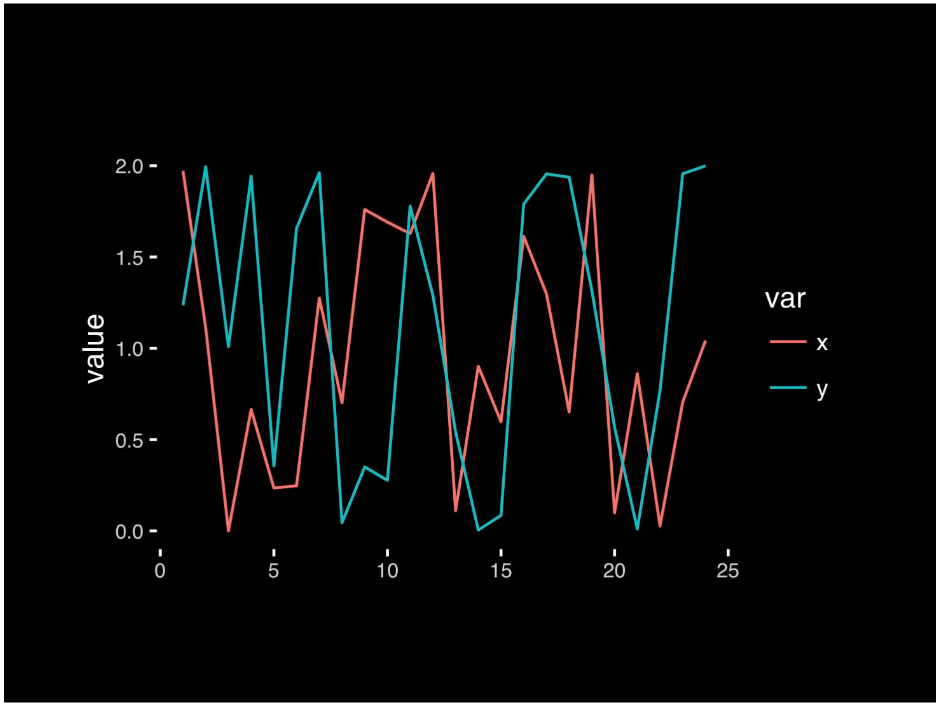
National Market System

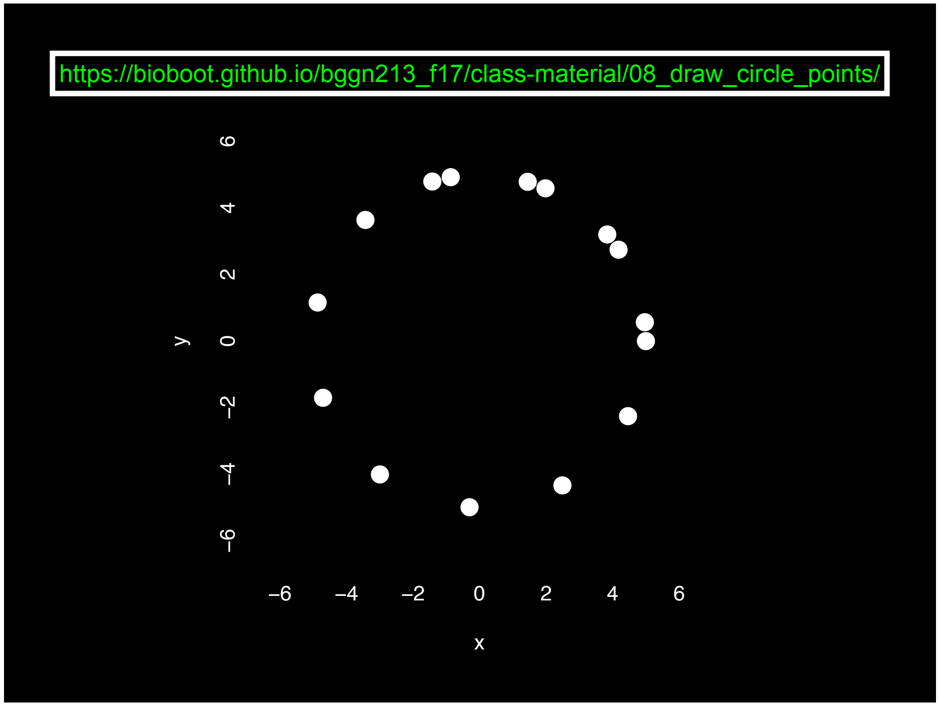
The companies listed below reflect the volume in 100's of shares on a daily basis and the closing price and net change are reflected from the previous day based on trades as provided under the NASD National Market System.

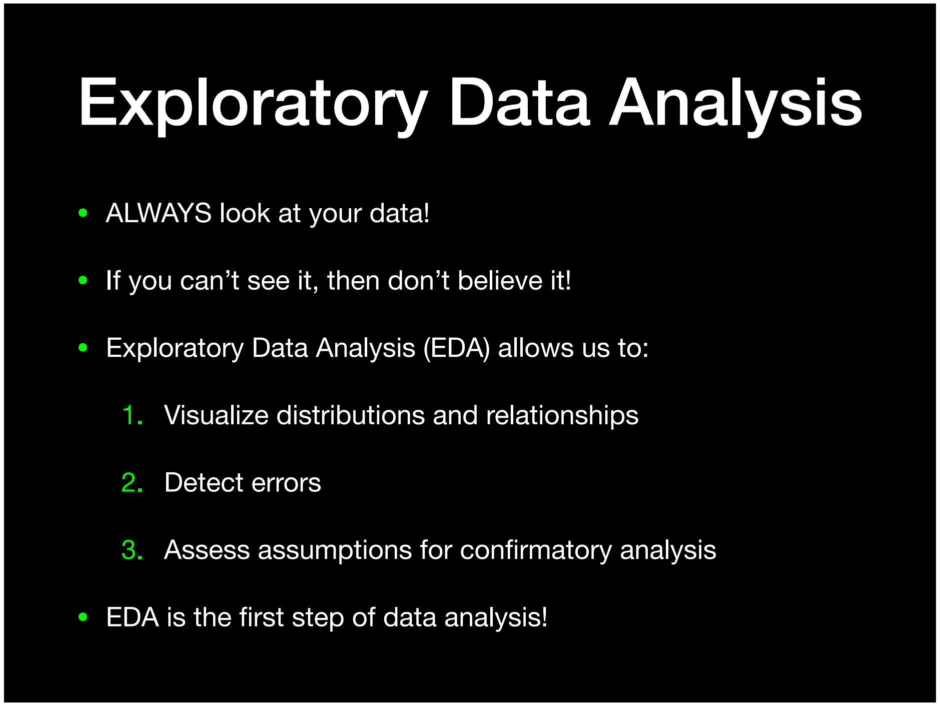
	x	y
1	5.00	0.00
2	4.18	2.75
3	1.98	4.59
4	-0.86	4.92
5	-3.43	3.64
6	-4.86	1.16
7	-4.70	-1.70
8	-2.99	-4.01
9	-0.30	-4.99
10	2.49	-4.34
11	4.46	-2.25
12	4.97	0.57
13	3.84	3.20
14	1.45	4.79
15	-1.42	4.79

	x	y
Min.	-4.86	-4.99
1st Qu.	-2.21	-1.98
Median	1.45	1.16
Mean	0.65	0.87
3rd Qu.	4.01	4.12
Max.	5.00	4.92






https://bioboot.github.io/bggn213_f17/class-material/08_draw_circle_points/


Exploratory Data Analysis

- ALWAYS look at your data!
- If you can’t see it, then don’t believe it!
- Exploratory Data Analysis (EDA) allows us to:
  1. Visualize distributions and relationships
  2. Detect errors
  3. Assess assumptions for confirmatory analysis
- EDA is the first step of data analysis!

Exploratory Data Analysis 1977

- Based on insights developed at Bell Labs in the 60's
- Techniques for visualizing and summarizing data
- What can the data tell us? (in contrast to "confirmatory" data analysis)
- Introduced many basic techniques:
  - 5-number summary, box plots, stem and leaf diagrams,…
- 5 Number summary:
  - extremes (min and max)
  - median & quartiles
  - More robust to skewed & longtailed distributions

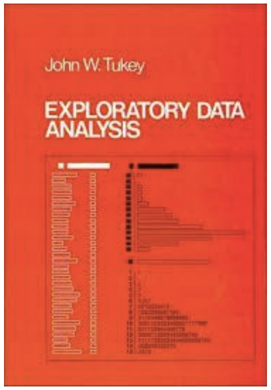
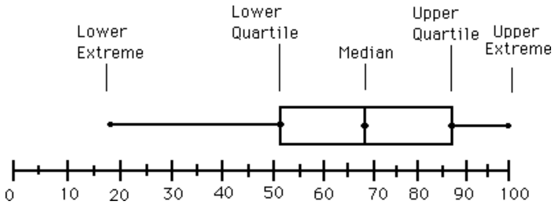
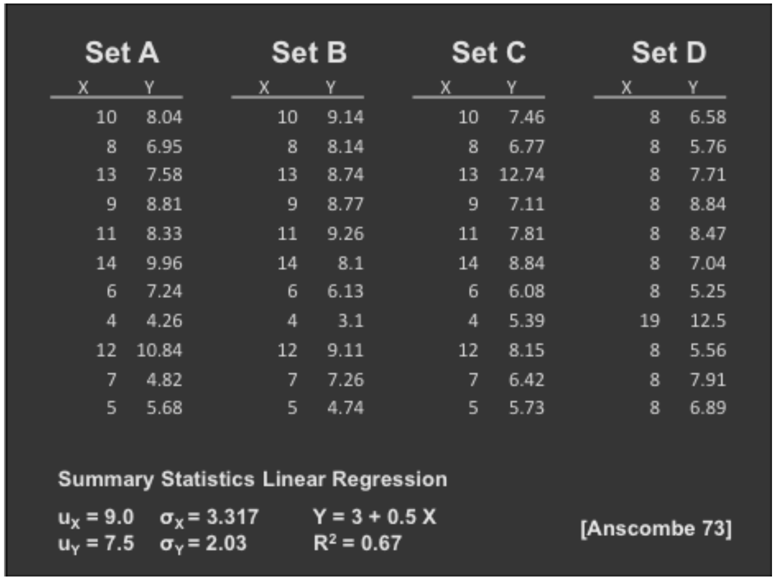


Chart types

- **Box-and-whisker plot** : a graphical form of 5-number summary (Tukey)



The Trouble with Summary Stats



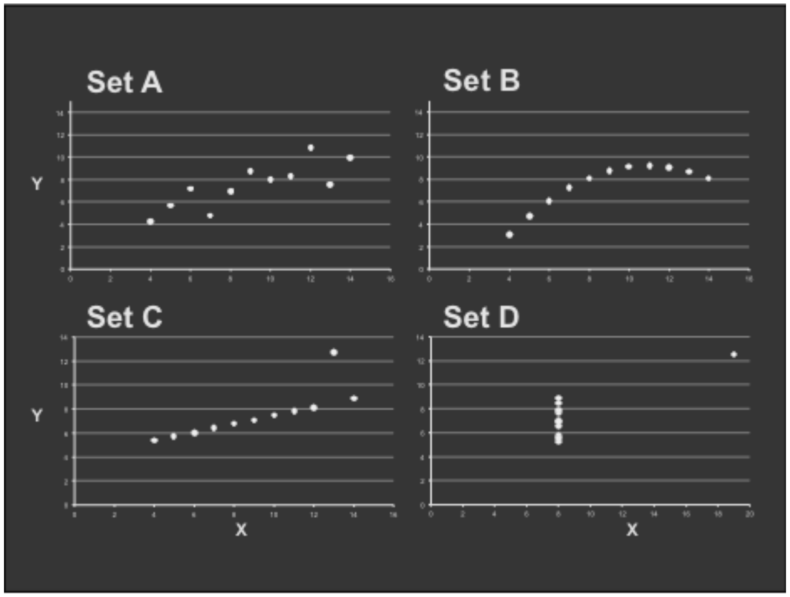
Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

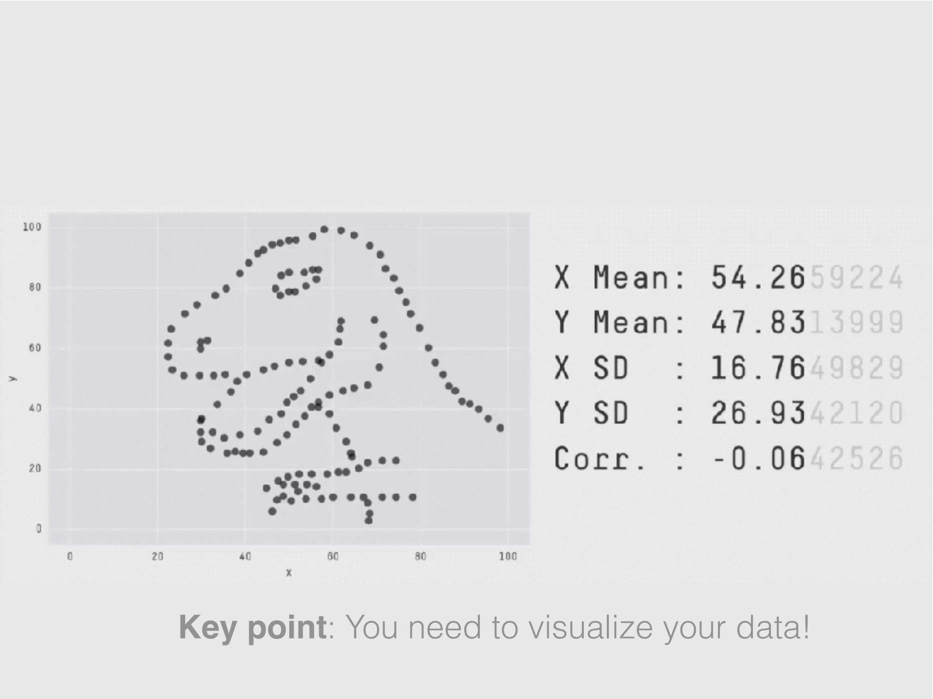
Summary Statistics Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\ X$
$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]

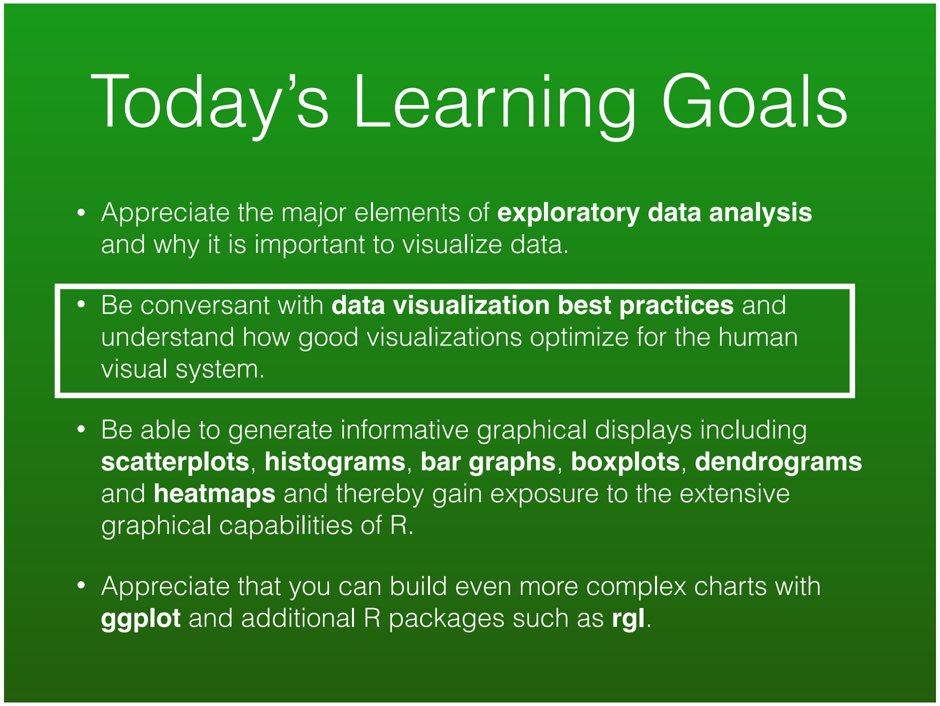
Looking at Data





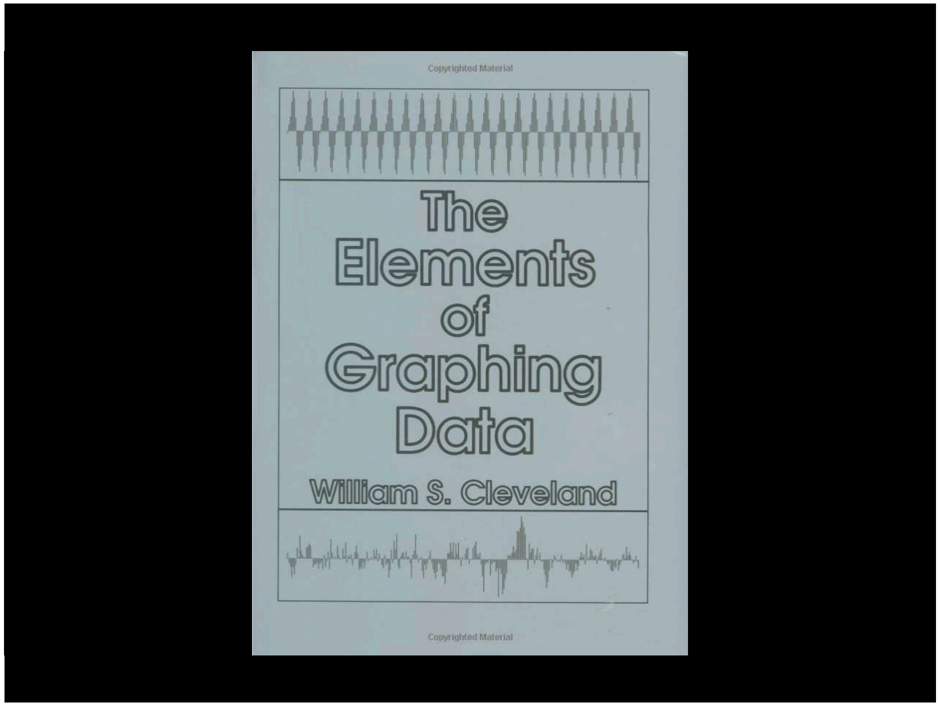
```
X Mean: 54.2659224
Y Mean: 47.8313999
X SD  : 16.7649829
Y SD  : 26.9342120
Corr. : -0.0642526
```

Key point: You need to visualize your data!



Today's Learning Goals

- Appreciate the major elements of **exploratory data analysis** and why it is important to visualize data.
- Be conversant with **data visualization best practices** and understand how good visualizations optimize for the human visual system.
- Be able to generate informative graphical displays including **scatterplots**, **histograms**, **bar graphs**, **boxplots**, **dendrograms** and **heatmaps** and thereby gain exposure to the extensive graphical capabilities of R.
- Appreciate that you can build even more complex charts with **ggplot** and additional R packages such as **rgl**.

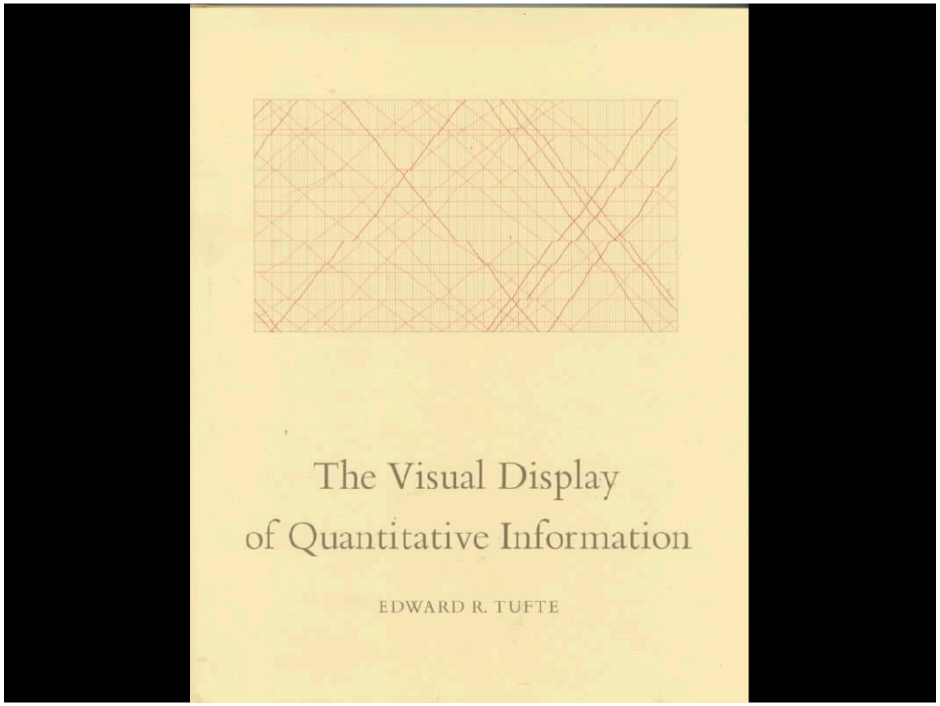


Copyrighted Material

The Elements of Graphing Data

William S. Cleveland

Copyrighted Material



The Visual Display of Quantitative Information

EDWARD R. TUFTE

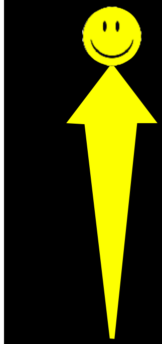
Key Point:

Good visualizations optimize for the human visual system.

Key Point: The most important measurement should exploit the highest ranked encoding possible

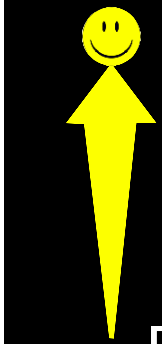
- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or Color saturation/hue

Key Point: The most important measurement should exploit the highest ranked encoding possible

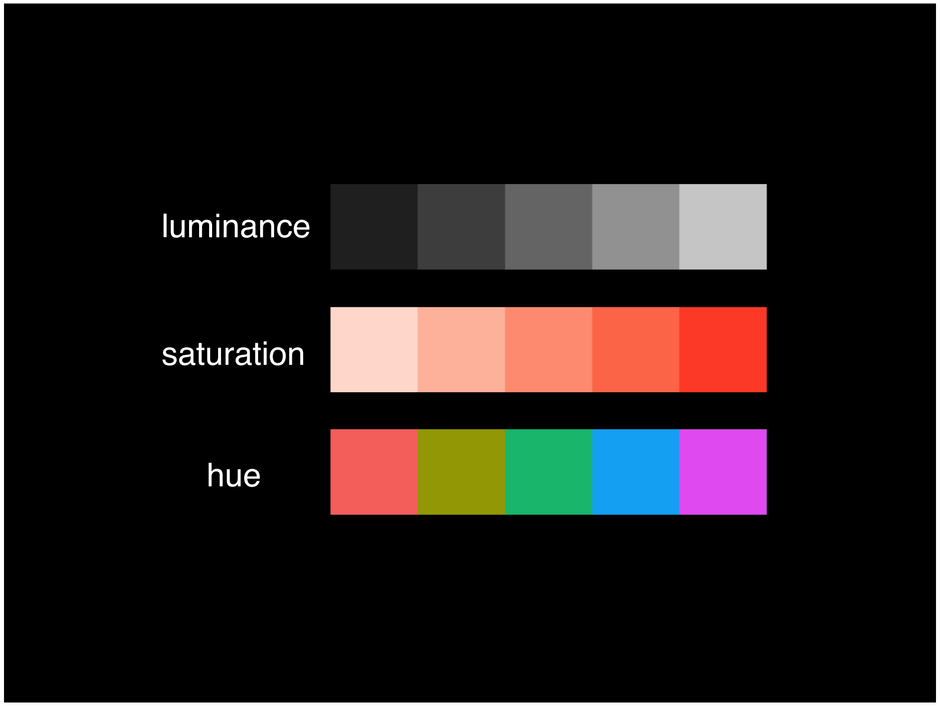


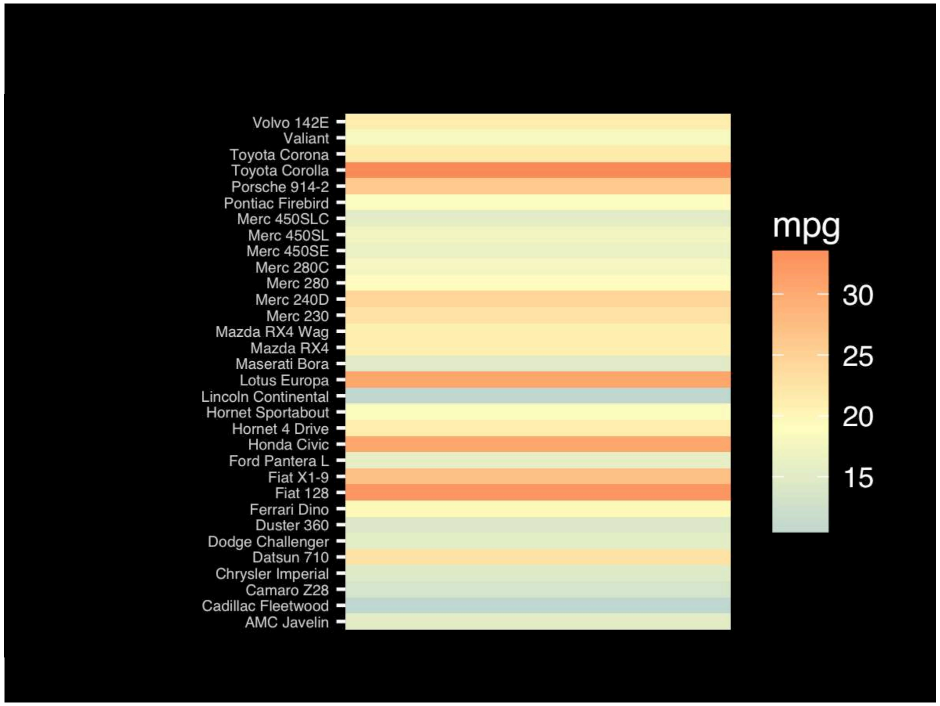
- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or Color saturation/hue

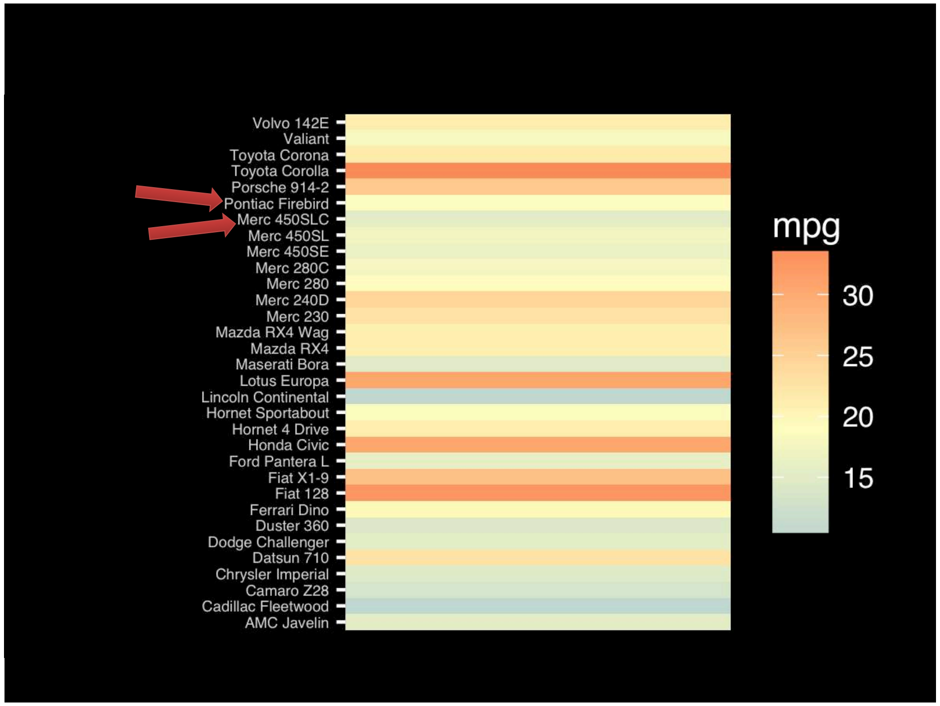
Key Point: The most important measurement should exploit the highest ranked encoding possible

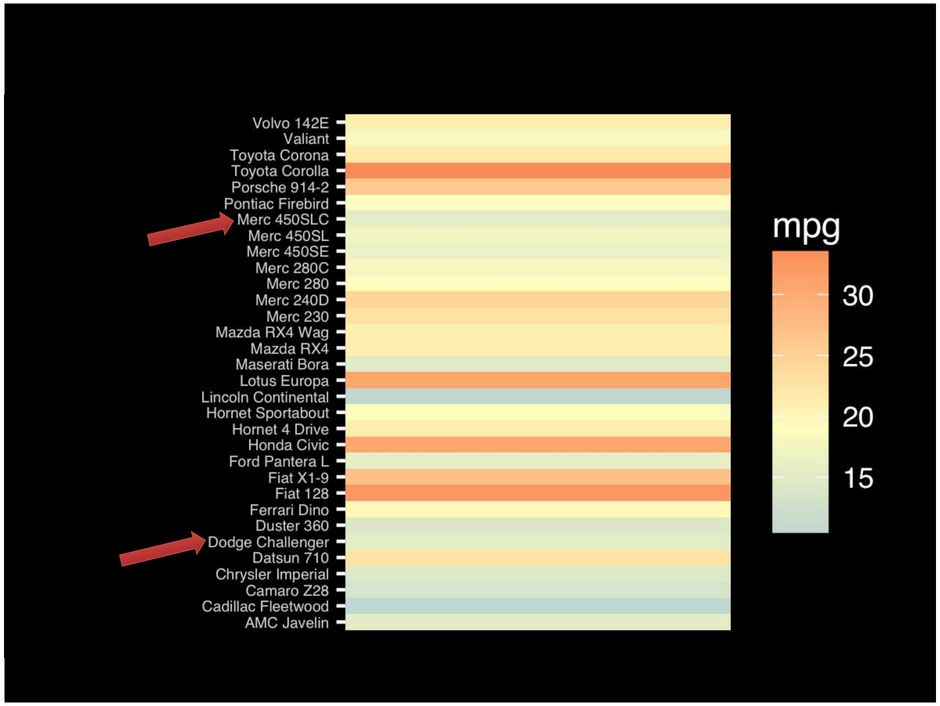


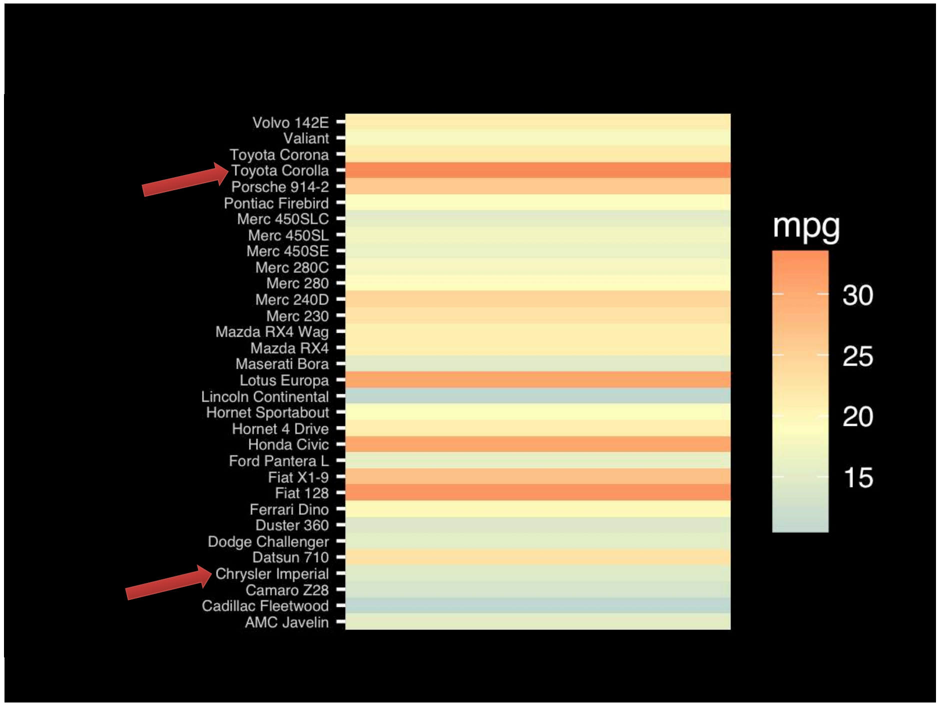
- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or **Color saturation/hue**

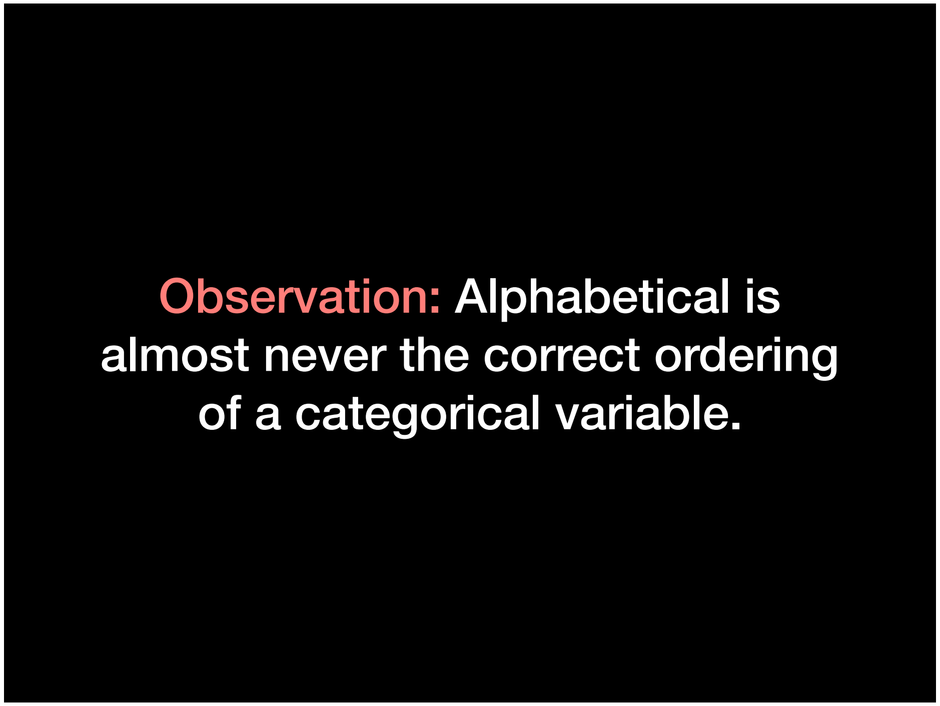

luminance
saturation
hue


Volvo 142E
Valiant
Toyota Corona
Toyota Corolla
Porsche 914-2
Pontiac Firebird
Merc 450SLC
Merc 450SL
Merc 450SE
Merc 280C
Merc 280
Merc 240D
Merc 230
Mazda RX4 Wag
Mazda RX4
Maserati Bora
Lotus Europa
Lincoln Continental
Hornet Sportabout
Hornet 4 Drive
Honda Civic
Ford Pantera L
Fiat X1-9
Fiat 128
Ferrari Dino
Duster 360
Dodge Challenger
Datsun 710
Chrysler Imperial
Camaro Z28
Cadillac Fleetwood
AMC Javelin
mpg
30
25
20
15

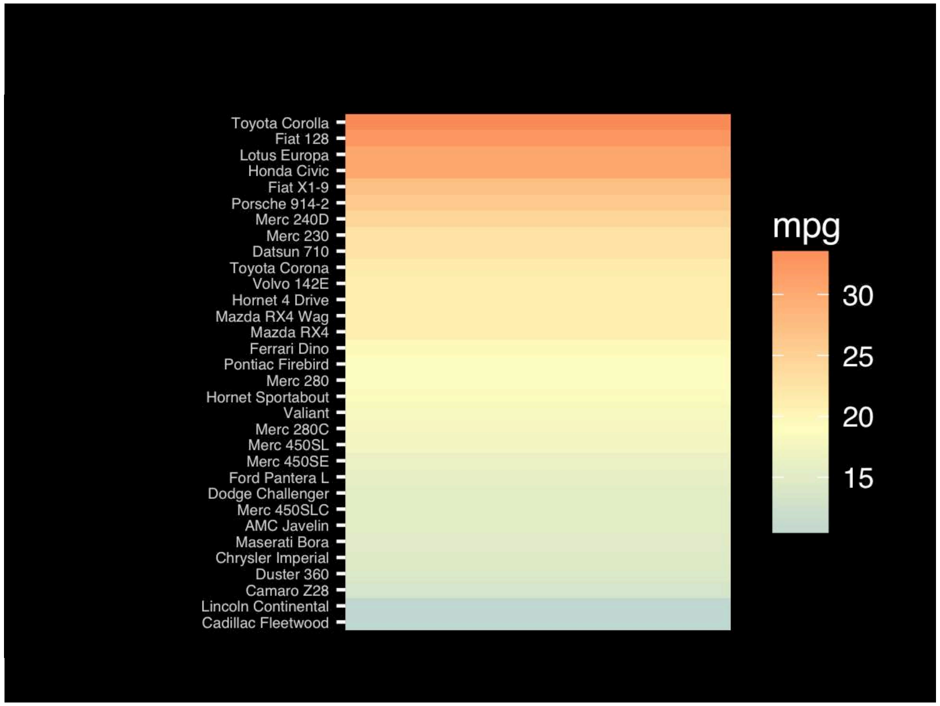

Volvo 142E
Valiant
Toyota Corona
Toyota Corolla
Porsche 914-2
Pontiac Firebird
Merc 450SLC
Merc 450SL
Merc 450SE
Merc 280C
Merc 280
Merc 240D
Merc 230
Mazda RX4 Wag
Mazda RX4
Maserati Bora
Lotus Europa
Lincoln Continental
Hornet Sportabout
Hornet 4 Drive
Honda Civic
Ford Pantera L
Fiat X1-9
Fiat 128
Ferrari Dino
Duster 360
Dodge Challenger
Datsun 710
Chrysler Imperial
Camaro Z28
Cadillac Fleetwood
AMC Javelin
mpg
30
25
20
15

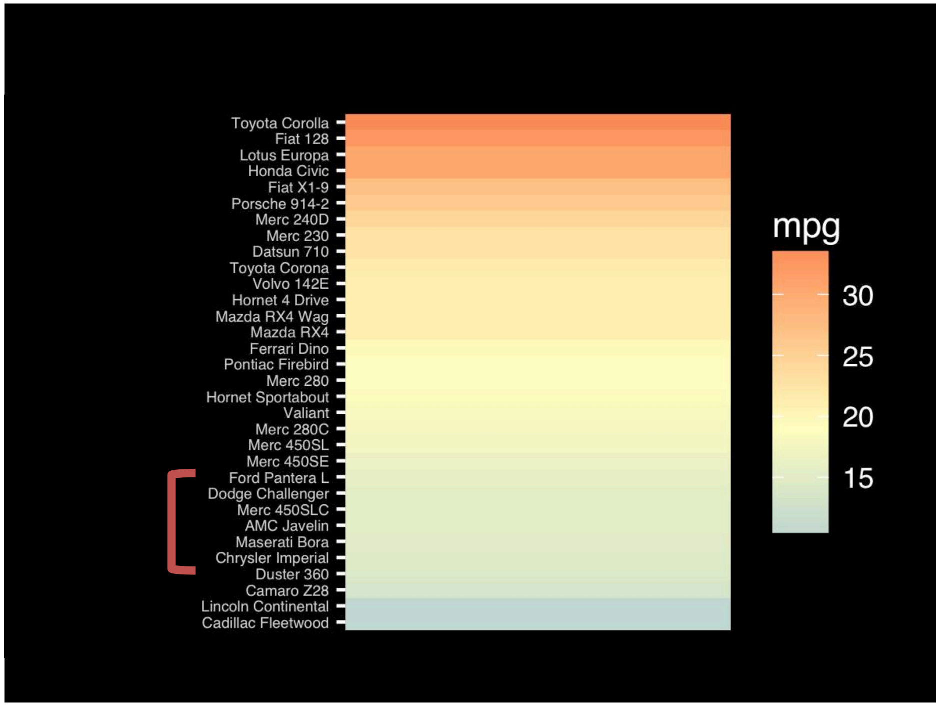

Volvo 142E
Valiant
Toyota Corona
Toyota Corolla
Porsche 914-2
Pontiac Firebird
Merc 450SLC
Merc 450SL
Merc 450SE
Merc 280C
Merc 280
Merc 240D
Merc 230
Mazda RX4 Wag
Mazda RX4
Maserati Bora
Lotus Europa
Lincoln Continental
Hornet Sportabout
Hornet 4 Drive
Honda Civic
Ford Pantera L
Fiat X1-9
Fiat 128
Ferrari Dino
Duster 360
Dodge Challenger
Datsun 710
Chrysler Imperial
Camaro Z28
Cadillac Fleetwood
AMC Javelin
mpg
30
25
20
15

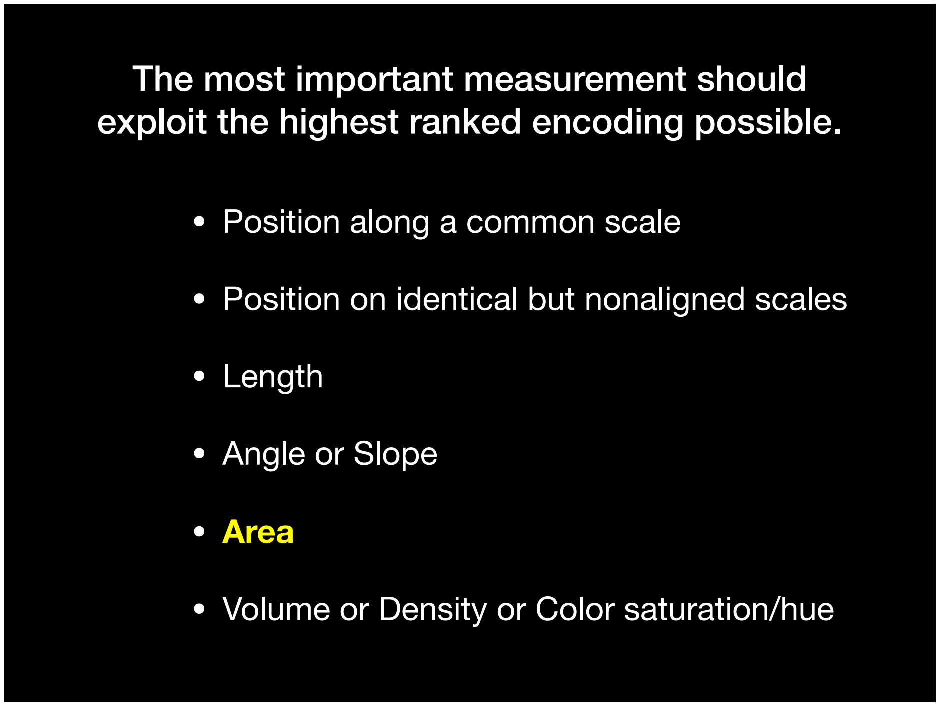




Observation: Alphabetical is almost never the correct ordering of a categorical variable.

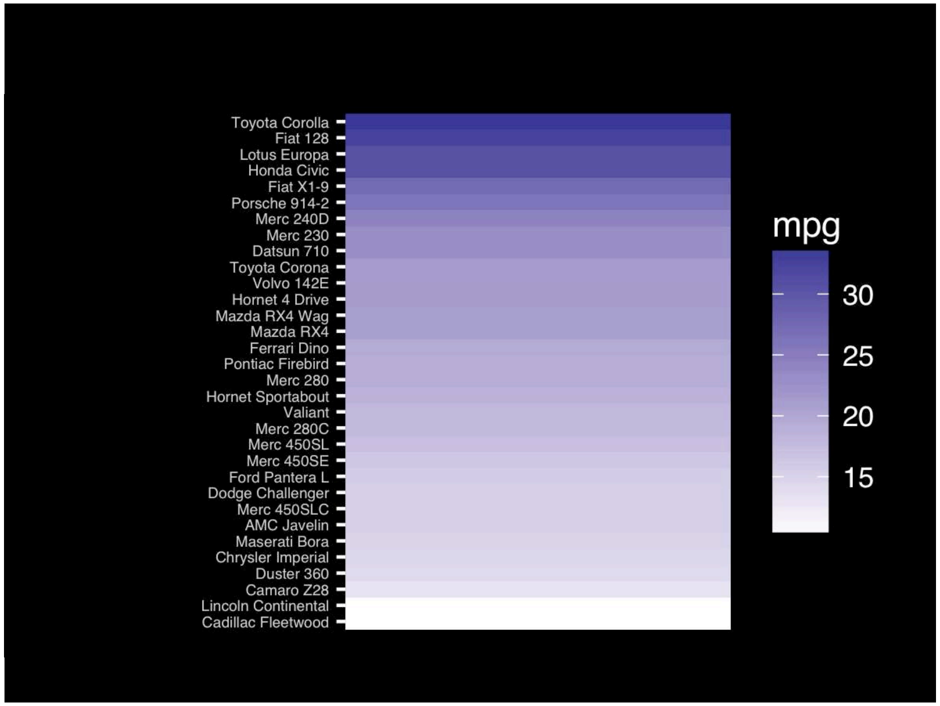


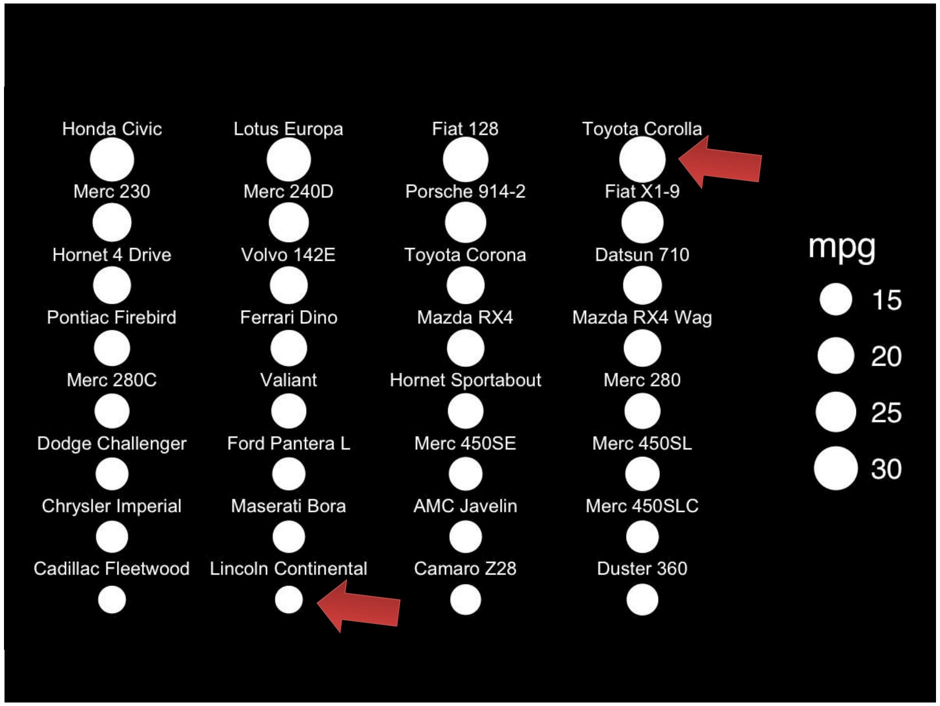


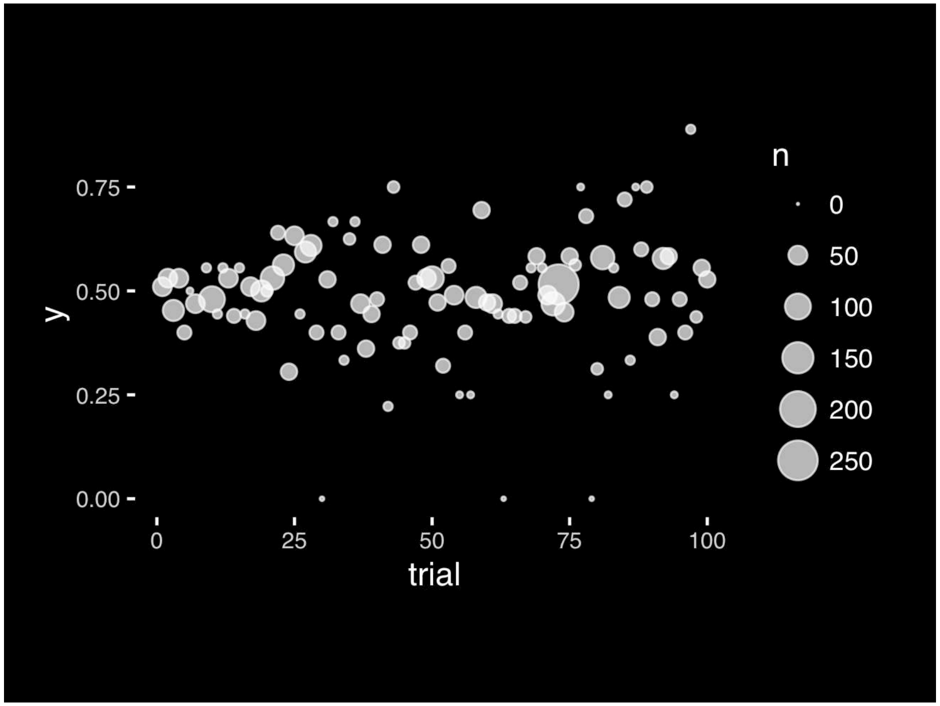


The most important measurement should exploit the highest ranked encoding possible.

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- **Area**
- Volume or Density or Color saturation/hue

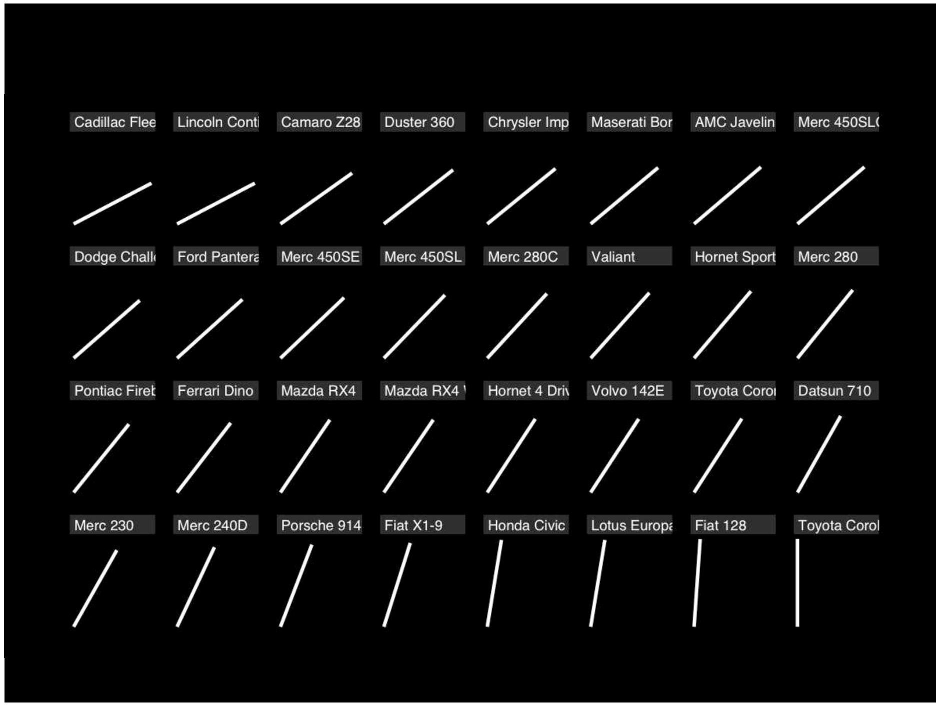


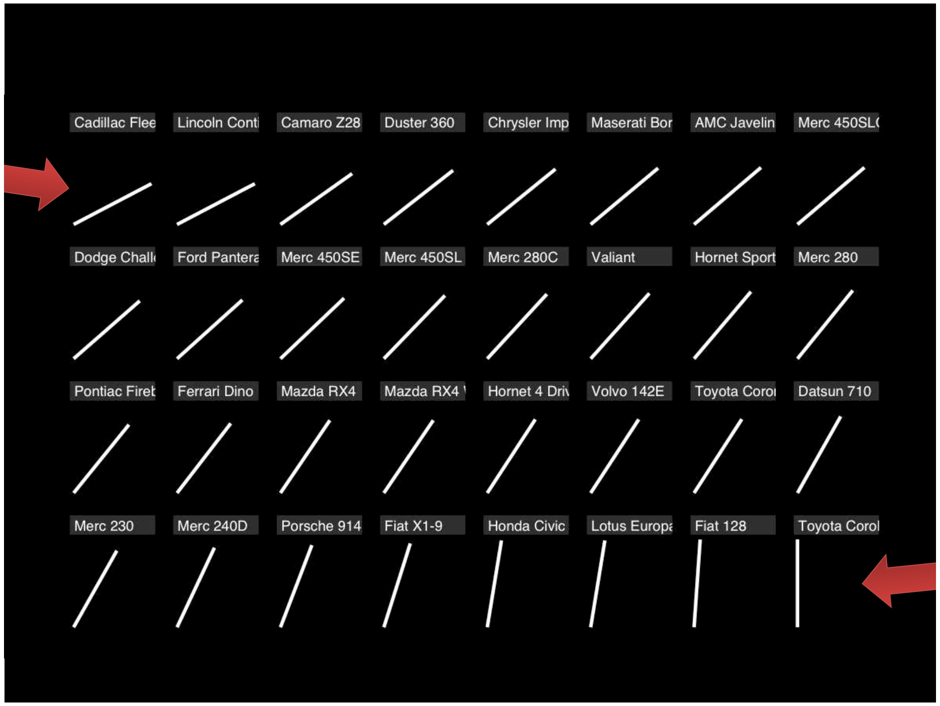


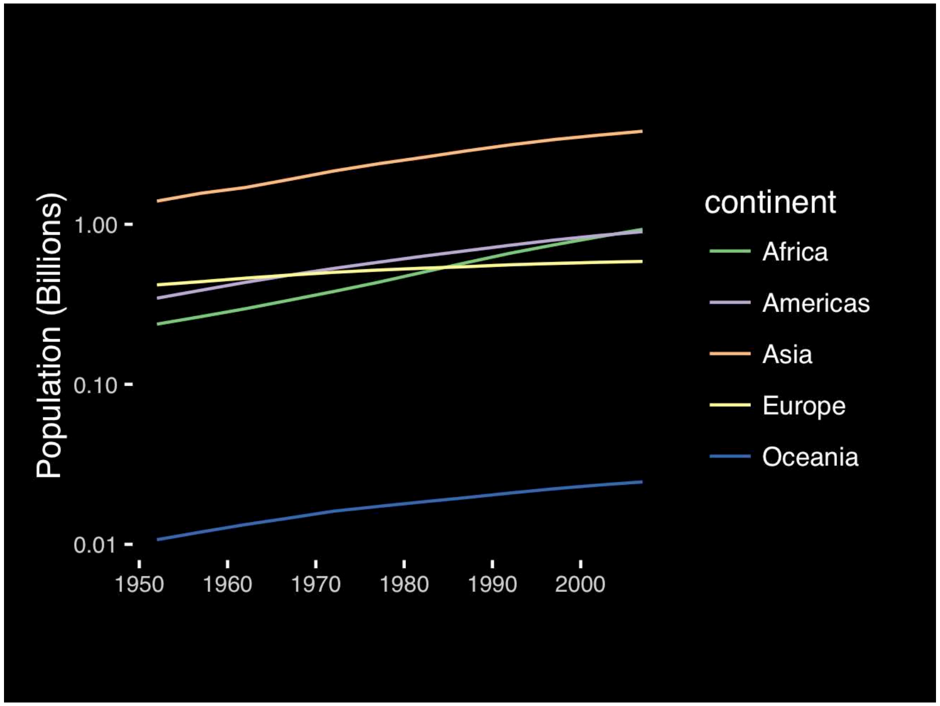


The most important measurement should exploit the highest ranked encoding possible.

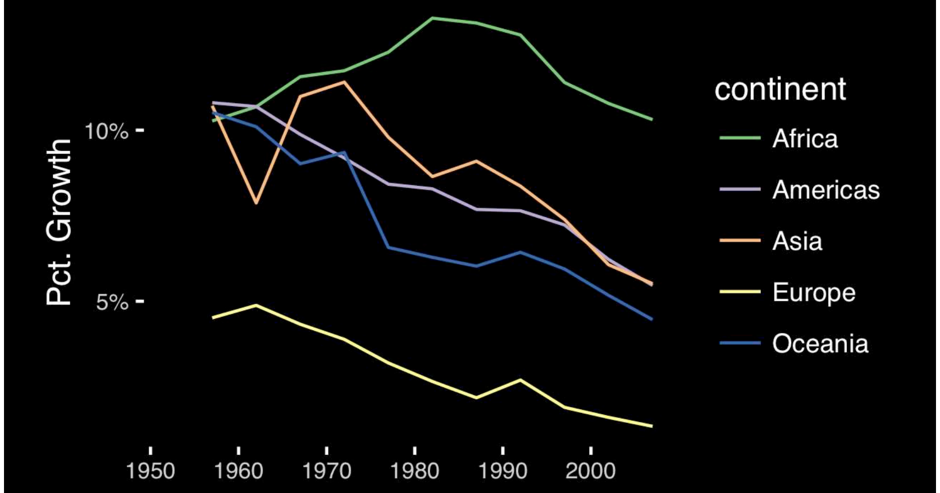
- Position along a common scale
- Position on identical but nonaligned scales
- Length
- **Angle or Slope**
- Area
- Volume or Density or Color saturation/hue







If growth (slope) is important, plot it directly.



The most important measurement should exploit the highest ranked encoding possible.

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- **Angle or Slope**
- Area
- Volume or Density or Color saturation/hue

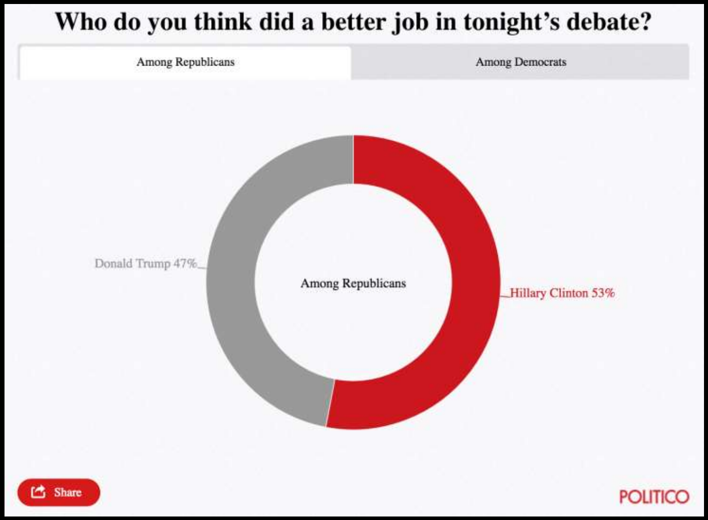
Observation: Pie charts are ALWAYS a mistake.

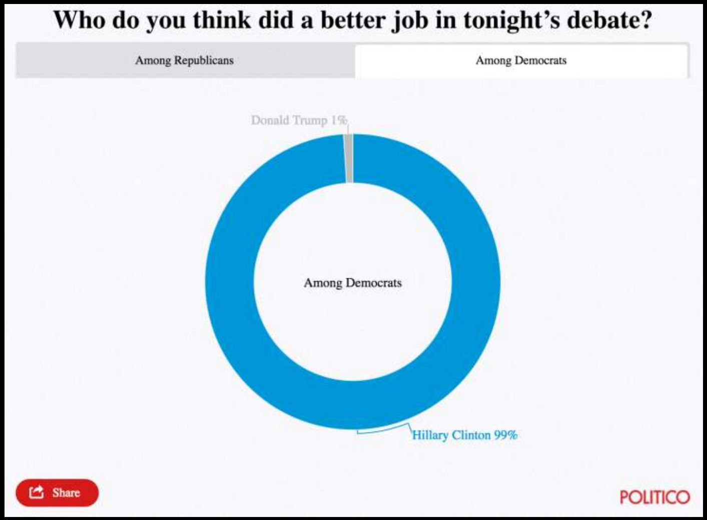
Piecharts are the information visualization equivalent of a roofing hammer to the frontal lobe. They have no place in the world of grownups, and occupy the same semiotic space as short pants, a runny nose, and chocolate smeared on one's face. They are as professional as a pair of assless chaps.

http://blog.codahale.com/2006/04/29/google-analytics-the-goggles-they-do-nothing/

Piecharts are the information visualization equivalent of a roofing hammer to the frontal lobe. They have no place in the world of grownups, and occupy the same semiotic space as short pants, a runny nose, and chocolate smeared on one's face. **They are as professional as a pair of assless chaps.**

http://blog.codahale.com/2006/04/29/google-analytics-the-goggles-they-do-nothing/


Who do you think did a better job in tonight's debate?


Who do you think did a better job in tonight's debate?

Tables are preferable to graphics for many small data sets. A table is nearly always better than a dumb pie chart; the only thing worse than a pie chart is several of them, for then the viewer is asked to compared quantities located in spatial disarray both within and between pies… Given their low data-density and failure to order numbers along a visual dimension, **pie charts should never be used.**
.

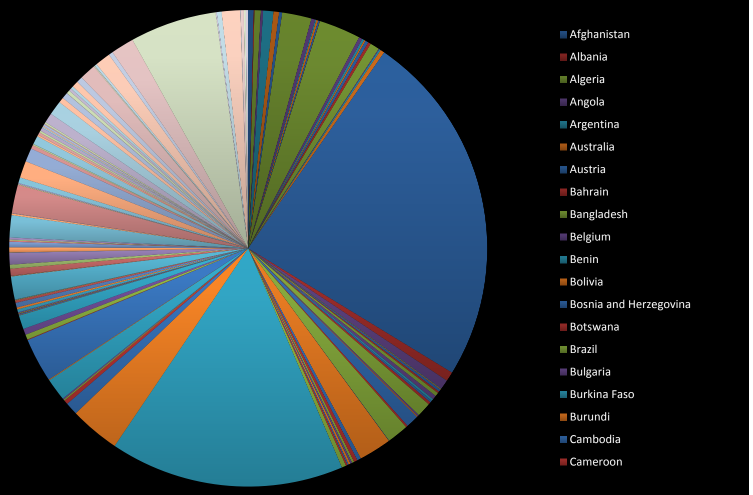
-Edward Tufte, The Visual Display of Quantitative Information

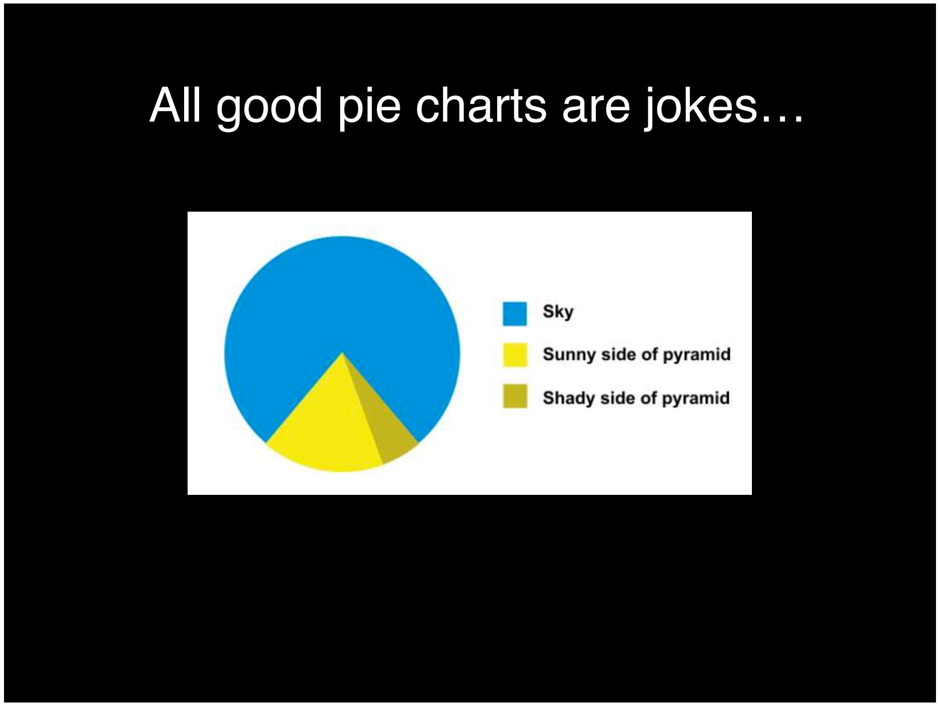
Tables are preferable to graphics for many small data sets. A table is nearly always better than a dumb pie chart; the only thing worse than a pie chart is several of them, for then the viewer is asked to compared quantities located in spatial disarray both within and between pies… Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.
.

-Edward Tufte, The Visual Display of Quantitative Information

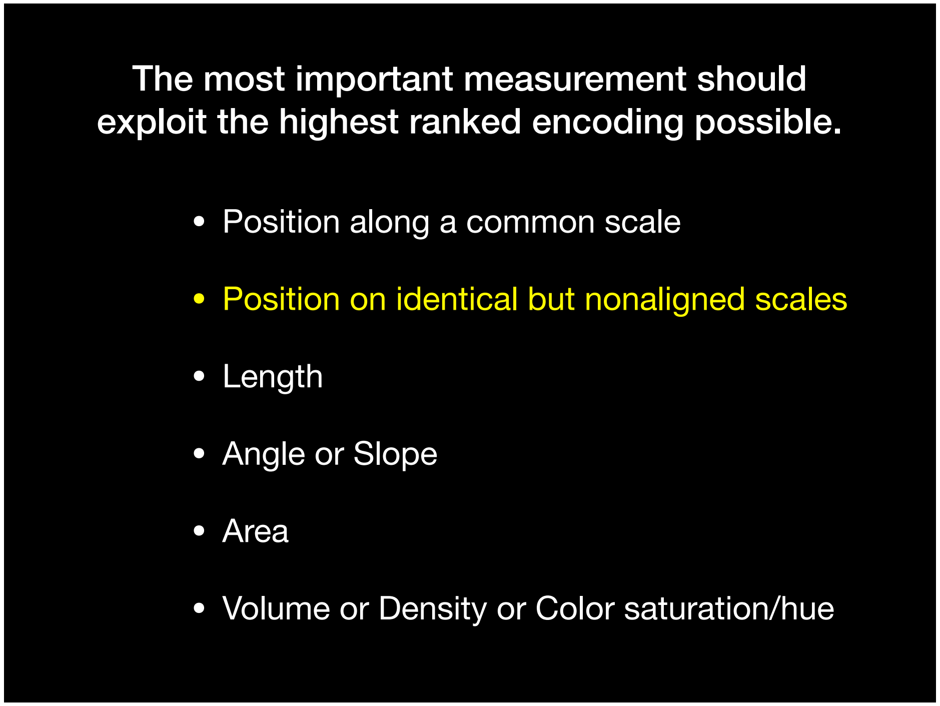
Who do you think did a better job in tonight's debate?

	Clinton	**Trump**
Among Democrats	99%	1%
Among Republicans	53%	47%



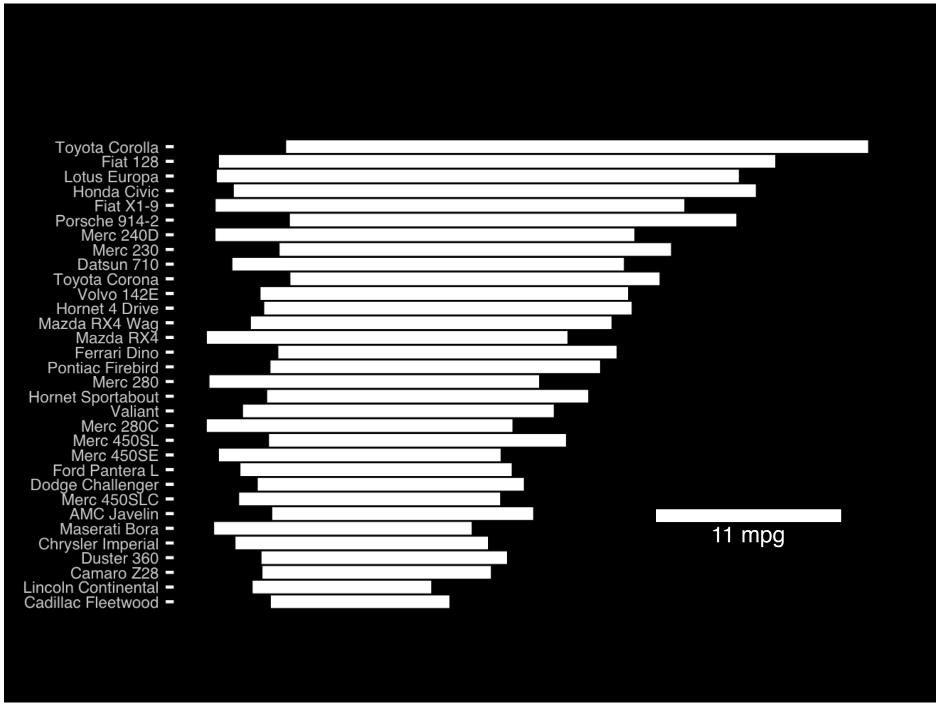


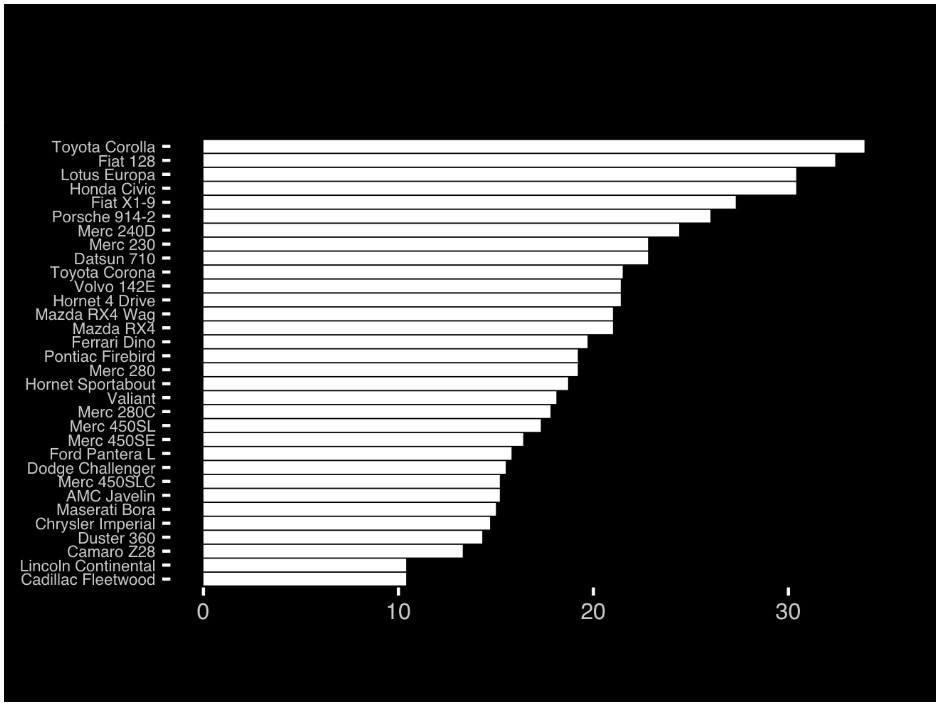
All good pie charts are jokes…

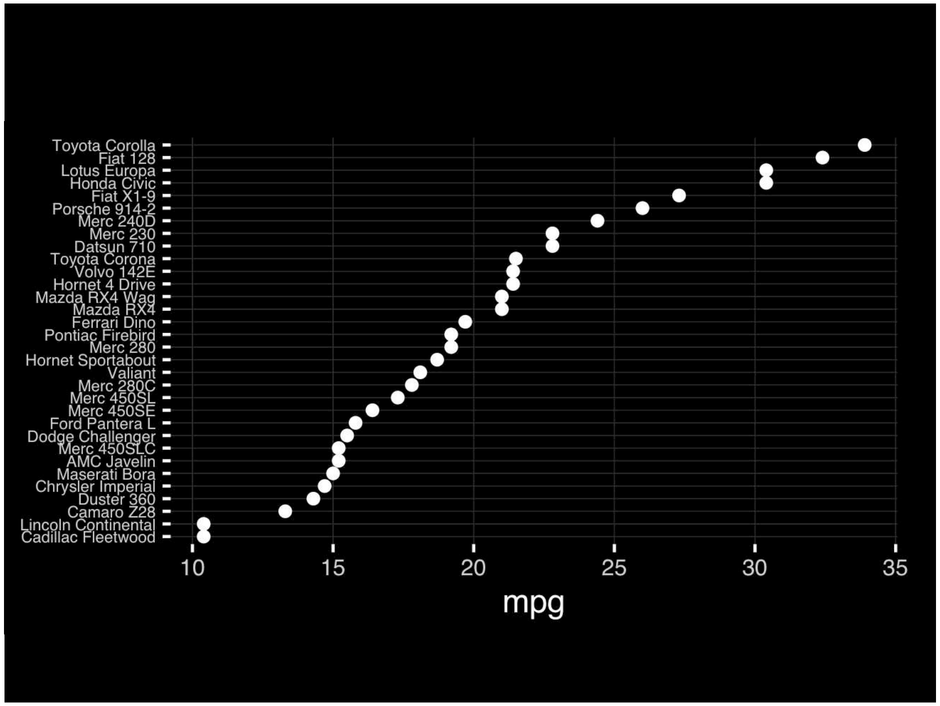


The most important measurement should exploit the highest ranked encoding possible.

- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or Color saturation/hue

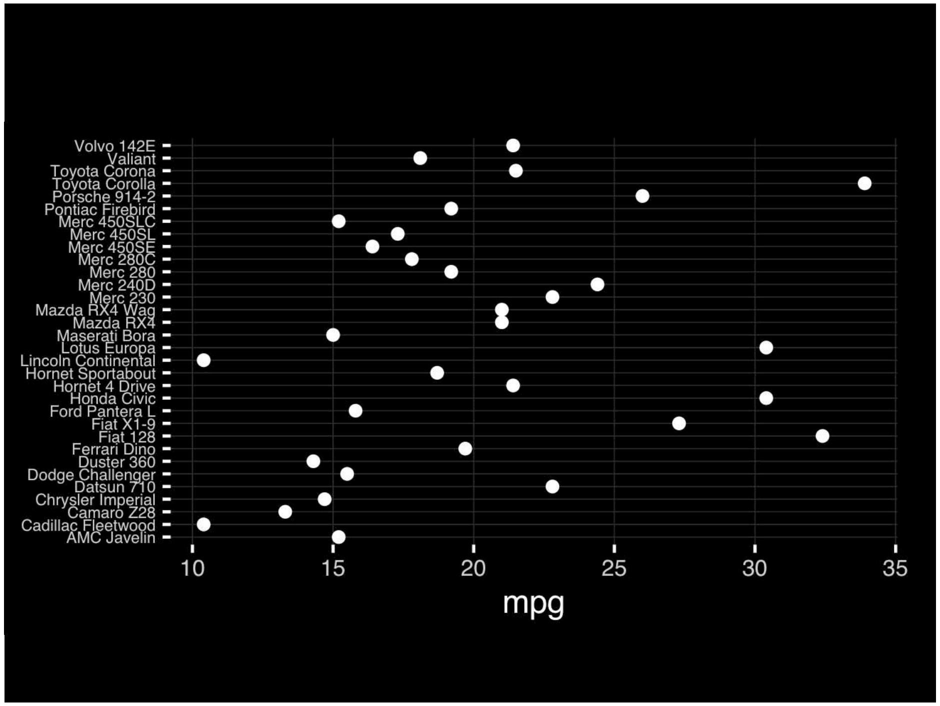


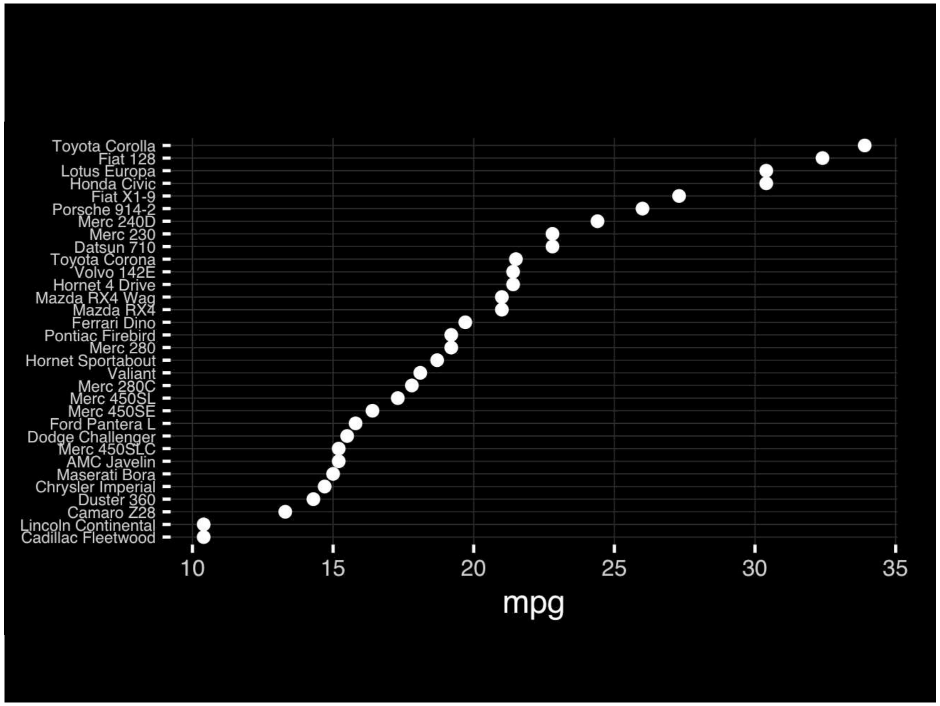


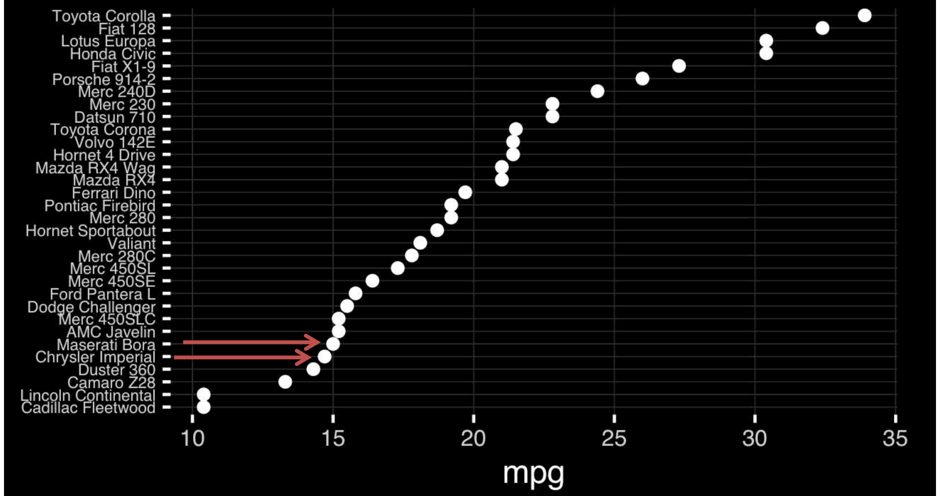


The most important measurement should exploit the highest ranked encoding possible.

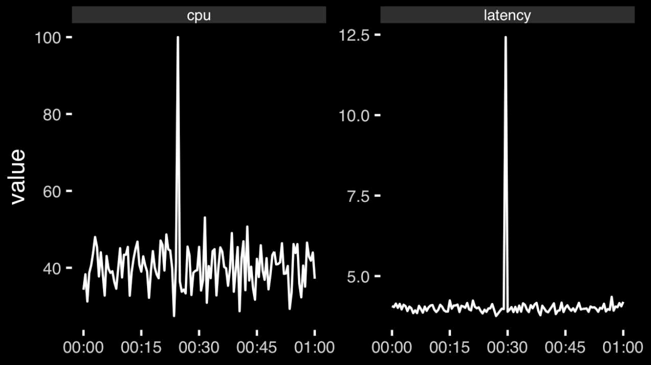
- Position along a common scale
- Position on identical but nonaligned scales
- Length
- Angle or Slope
- Area
- Volume or Density or Color saturation/hue

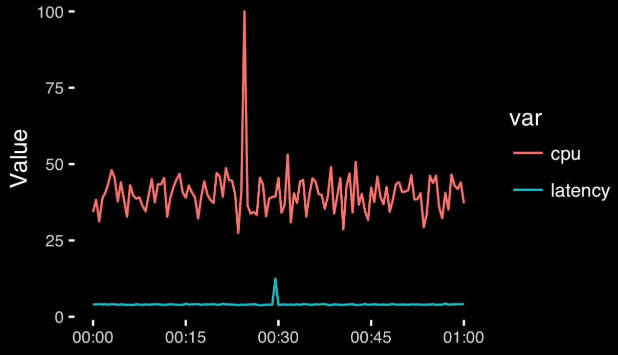


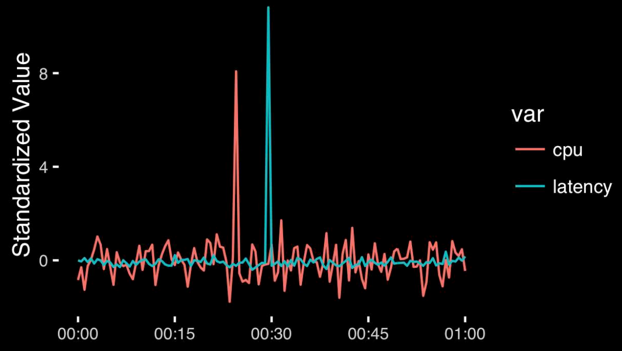




Observation: Comparison is trivial on a common scale.





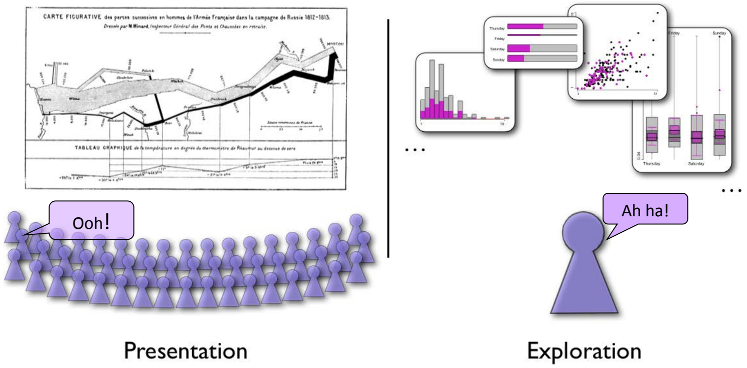


Today's Learning Goals

- Appreciate the major elements of **exploratory data analysis** and why it is important to visualize data.
- Be conversant with **data visualization best practices** and understand how good visualizations optimize for the human visual system.
- Be able to generate informative graphical displays including **scatterplots**, **histograms**, **bar graphs**, **boxplots**, **dendrograms** and **heatmaps** and thereby gain exposure to the extensive graphical capabilities of R.
- Appreciate that you can build even more complex charts with **ggplot** and additional R packages such as **rgl**.

Different graphs for different purposes

Exploratory graphs: many images for a narrow audience (you!)
Presentation graphs: single image for a large audience



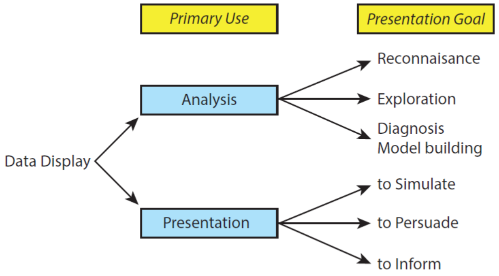
Roles of graphics in data analysis

- Graphs (& tables) are forms of communication:
  - What is the audience?
  - What is the message?

Analysis graphs: design to see patterns, trends, aid the process of data description, interpretation

Presentation graphs: design to attract attention, make a point, illustrate a conclusion

Basic functions of data display



Primary Use

Presentation Goal

Data Display

Analysis

Reconnaisance

Exploration

Diagnosis
Model building

Presentation

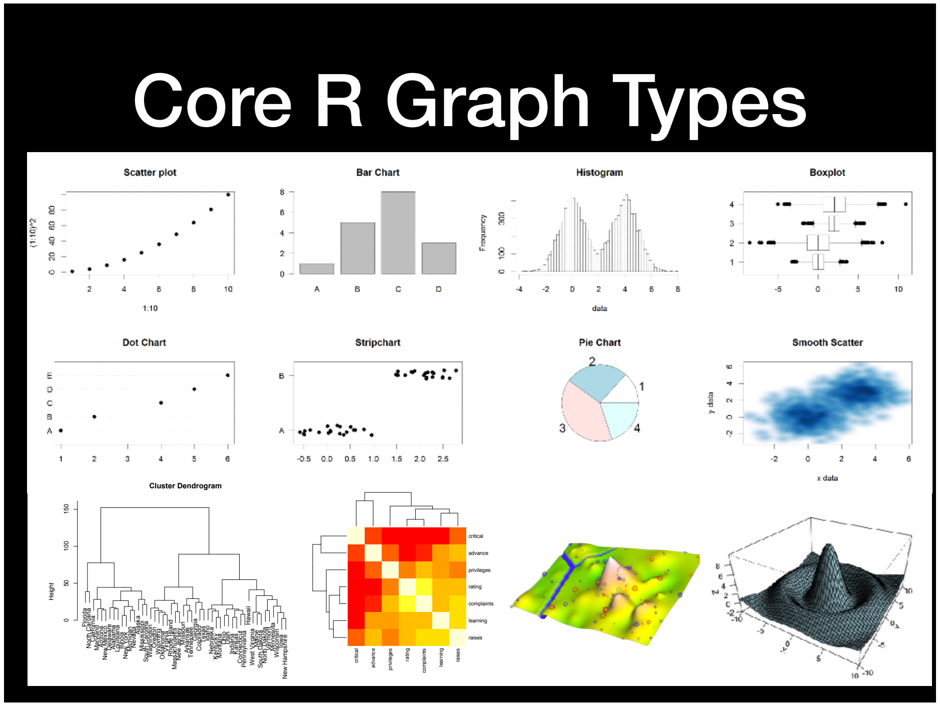
to Simulate

to Persuade

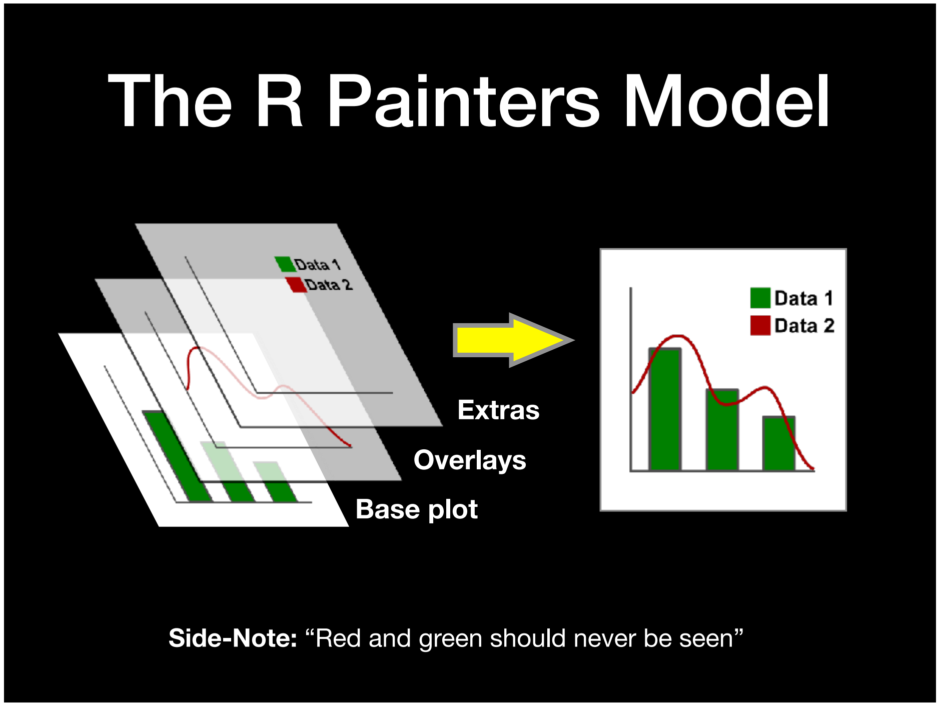
to Inform

Exploratory (for you!)

Info for others, publications & sharing etc.



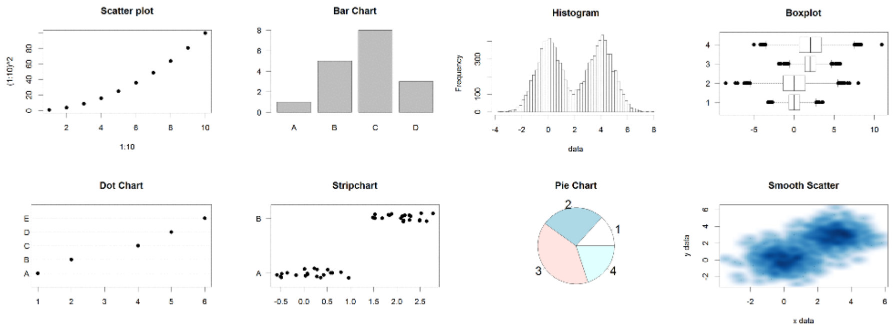
Core R Graph Types



The R Painters Model

Side-Note: "Red and green should never be seen"

Core Graph Types



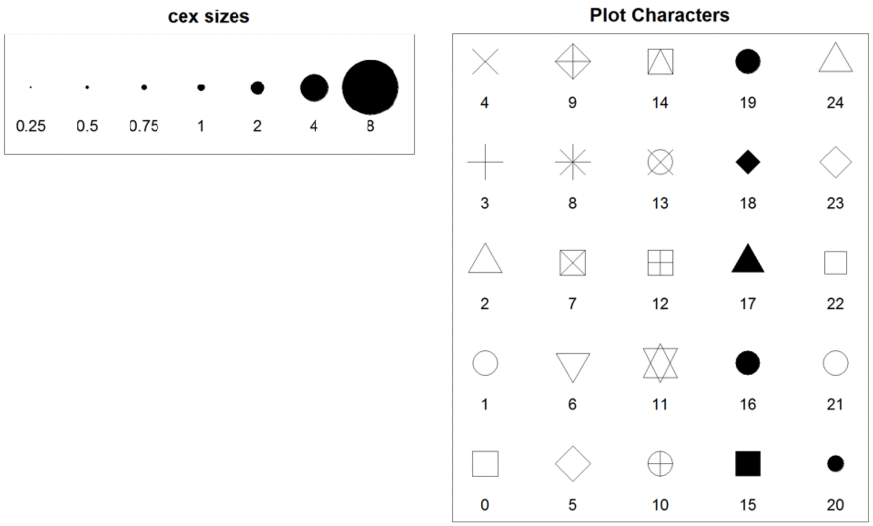
- Local options to change a specific plot
- Global options to affect all graphs

Common Options

- Axis scales
  - `xlim c(min,max)`
  - `ylim c(min,max)`
- Axis labels
  - `xlab(text)`
  - `ylab(text)`
- Plot titles
  - `main(text)`
  - `sub(text)`
- Plot characters
  - `pch(number)`
  - `cex(number)`

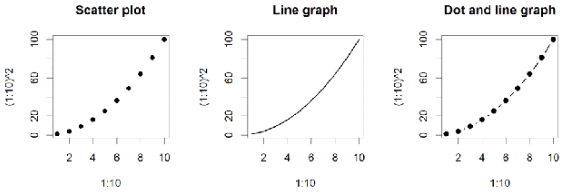
- Local options to change a specific plot
- Global options to affect all graphs

Plot Characters



Plot Type Specific Options

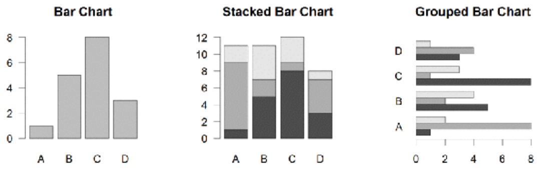
Plot (scatterplots and line graphs)



- Input: Almost anything. 2 x Vectors
- Output: Nothing
- Options:
  - `type` l=line, p=point, b=line+point
  - `lwd` line width (thickness)
  - `lty` line type (1=solid,2=dashed,3=dotted etc.)

```
plot( c(1:10)^2, typ="b", lwd=4, lty=3 )
```

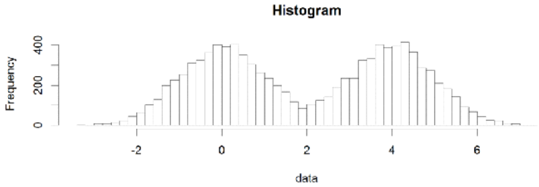
Barplot (bar graphs)



- Input: Vector (single) or Matrix (stack or group)
- Output: Bar centre positions
- Options:
  - `names.arg` Bar labels (if not from data)
  - `horiz=TRUE` Plot horizontally
  - `beside=TRUE` Plot multiple series as a group not stacked

```
barplot(VADeaths, beside = TRUE)
```

Hist (histograms)

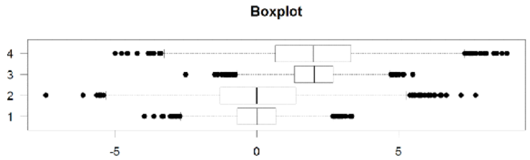


- Input: Vector
- Output: Summary of binned data
- Options:

– `breaks`	Number or limits of bins
– `probability`	Y axis is probability, not freq
– `labels`	Per bin text labels

```
hist( c( rnorm(1000,0), rnorm(1000,4) ), breaks=20 )
```

Boxplot



- Input: Vector, List or formula (`data~factor`)
- Output: Summary of the boxplot parameters
- Options:

– `range`	Sensitivity of whiskers
– `varwidth`	Width represents total observations
– `horizontal`	Plot horizontally

```
boxplot( cbind( rnorm(1000,0), rnorm(1000,4) ) )
```

Controlling plot area options with `par`

Par

- The `par` function controls global parameters affecting all plots in the current plot area

- Changes affect all subsequent plots

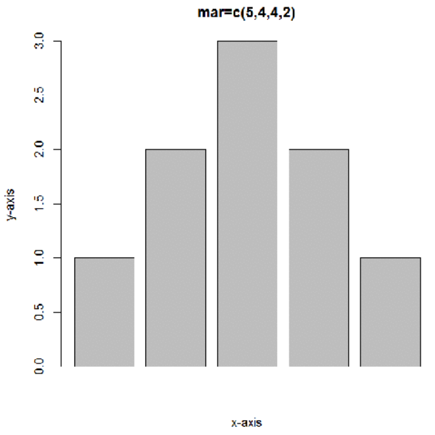
- Many par options can also be passed to individual plots

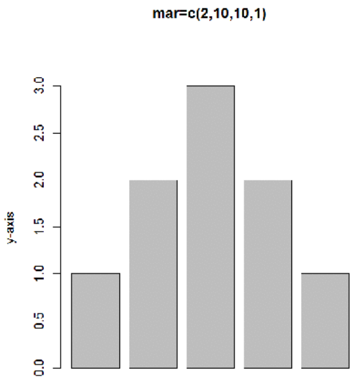
Par examples

- Reading current value
  - `par()$cex`
- Setting a value
  - `par(cex=1.5) -> old.par`
- Restoring a value
  - `par(old.par)`
  - `dev.off()`

Par options

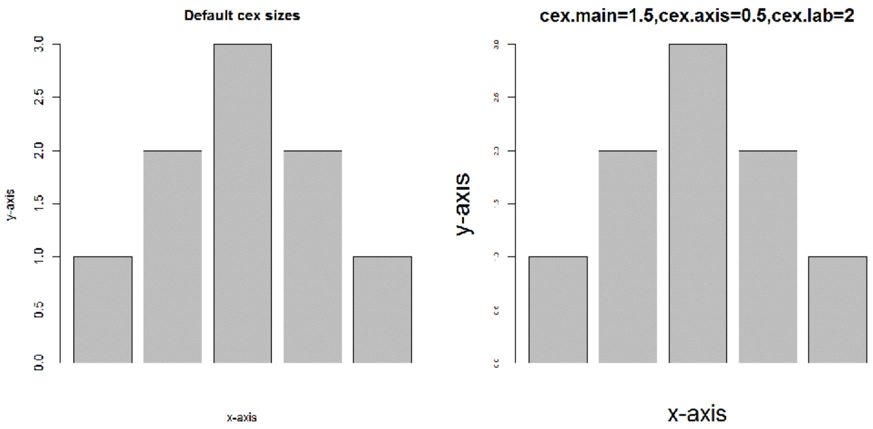
- Margins
  - `mai` (set margins in inches)
  - `mar` (set margins in number of lines)
  - `mex` (set lines per inch)
  - 4 element vector (bottom, left, top, right)
- Warning
  - `Error in plot.new() : figure margins too large`





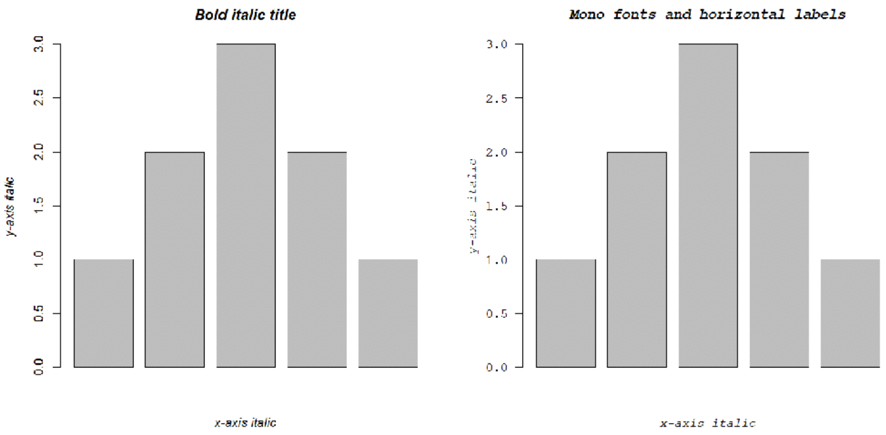
Par options

- Fonts and labels
  - `cex` - global char expansion
    - `cex.axis`
    - `cex.lab`
    - `cex.main`
    - `cex.sub`



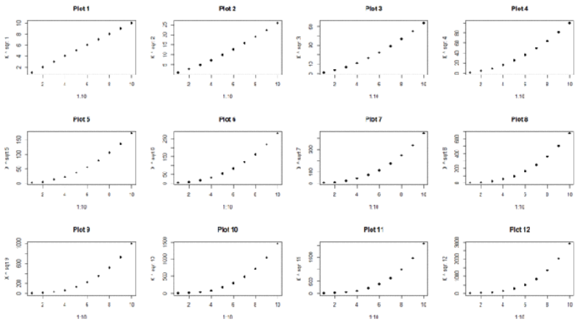
Par options

- Font style
  - `font` (`font.axis,font.main,font.sub,font.lab`)
    - 1 = Plain text
    - 2 = Bold text
    - 3 = Italic text
    - 4 = Bold italic text
  - `las` (label orientation)
    - 0 = Parallel to axis
    - 1 = Horizontal
    - 2 = Perpendicular
    - 3 = Vertical



Par options

- Multi-panel
  - `mfrow(rows,cols)`
  - Not supported by some packages



Exercise 1

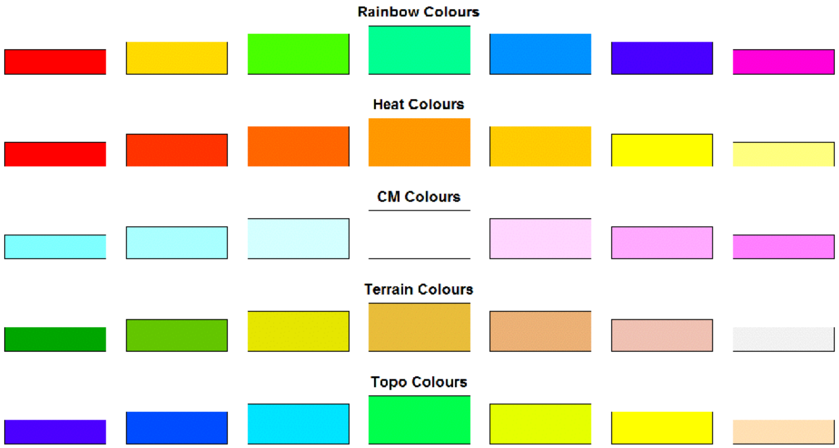
Using Color

Specifying colors

- Hexadecimal strings
  - `#FF0000` (red)
  - `#0000FF` (blue)
  - `#CC00CC` (purple)
- Controlled names
  - `“red” “green”` etc.
  - `colors()`

Built in color schemes

- Functions to generate colors
- Pass in number of colors to make
- Functions:
  - `rainbow`
  - `heat.colors`
  - `cm.colors`
  - `terrain.colors`
  - `topo.colors`

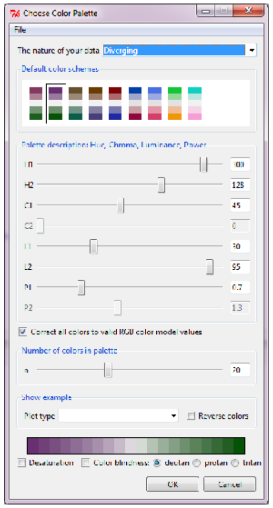


Color Packages

- Color Brewer
  - Set of pre-defined, optimized palettes
  - `library(RColorBrewer)`
  - `brewer.pal(no colours, palette)`
- ColorRamps
  - Create smooth palettes for ramped color
  - Generates a function to make actual color vectors
  - `colorRampPalette(c(“red”,”white”,”blue”))`
  - `colorRampPalette(c(“red”,”white”,”blue”))(5)`

Color Packages

- Colorspace
  - `library(colorspace)`
  - `choose.palette()`


Choose Color Palette
File
Default color schemes
Palette description: Hue, Chroma, Luminance, Power
Number of colors in palette
Show example
OK
Cancel

The nature of your data: Diverging

H1: 300
H2: 128
C1: 45
C2: 0
L1: 30
L2: 95
P1: 0.7
P2: 1.3

☑ Correct all colors to valid RGB color model values

n: 20

Plot type:
□ Reverse colors

□ Desaturation □ Color blindness: ● deutan ○ protan ○ tritan

Applying Color to Plots

- Vector of colors passed to the `col` parameter
- Vector of factors used to divide the data
  - Colors taken from pallete
  - Can read or set using pallete function
    - `palette()`
    - `palette(brewer.pal(9,”Set1”))`
    - Ordered by levels of factor vector

Dynamic use of color

- Coloring by density
  - Pass data and palette to `densCols`
  - Vector of colors returned
- Coloring by value
  - Need function to map values to colors

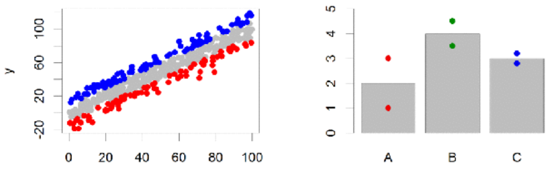
Color Mapping Function

```
map.colors <- function(value,range,palette) {

  proportion <- (value-range[1])/(range[2]-range[1])
  index <- round((length(palette)-1)*proportion)+1

  return(palette[index])
}
```

Exercise 2

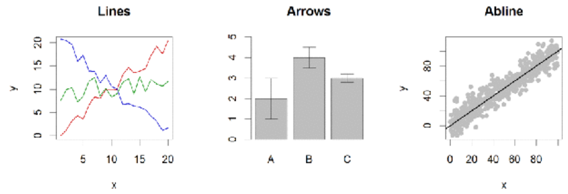
Plot Overlays
Exercise 3

Points



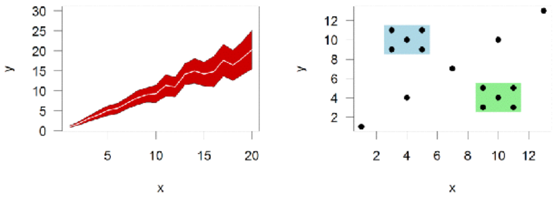
- Input: 2 Vectors (x and y positions)
- Options:
  - pch
  - cex

Lines / Arrows / Abline



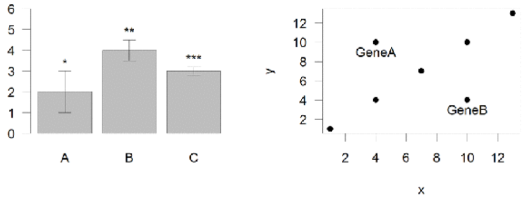
- Input:
  - Lines 2 vectors (x and y)
  - Arrows 4 vectors (x0,x1,y0,y1)
  - Abline Intercept and slope (or correlation object)
- Options:
  - lwd
  - angle (arrows)

Polygon (shaded areas)



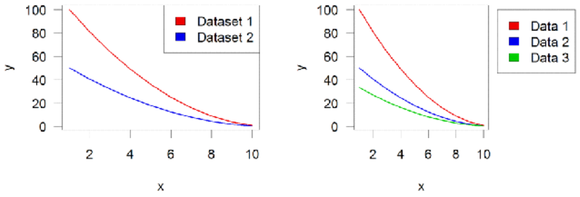
- Input:
  - 2 vectors (x and y) for bounding region
- Options:
  - col

Text (in plot text)



- Input:
  - Text, x, y
- Options:
  - adj (x and y offsets)
  - pos (auto offset 1=below,2=left,3=above, 4=right)

Legend



- Input:
  - Position (x,y or “topright”,”bottomleft” etc)
  - Text labels
- Options:
  - `fill` (colours for shaded boxes)
  - `xpd=NA` (draw outside plot area)

Exercise 3