DATA11001

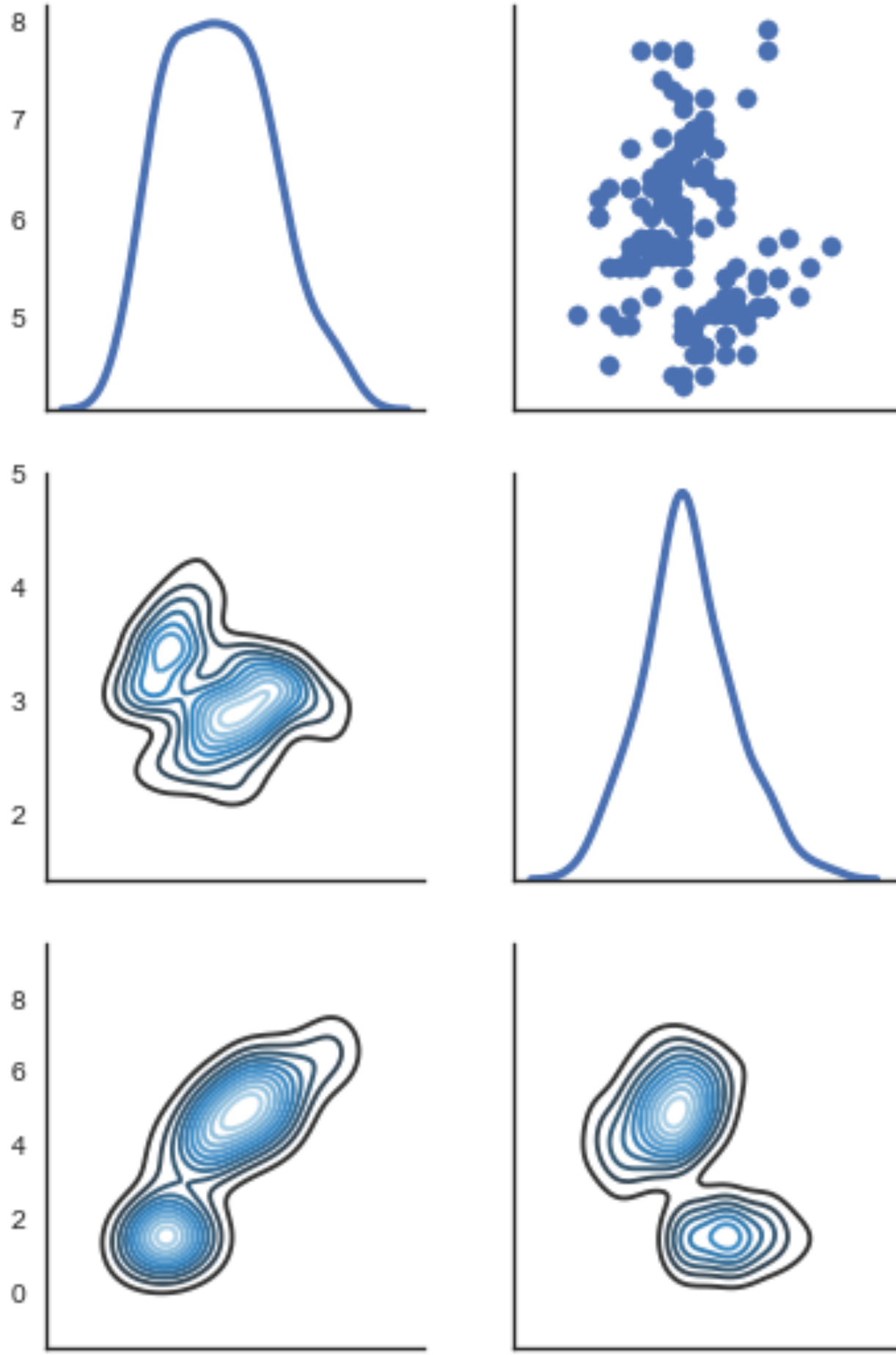
INTRODUCTION TO DATA SCIENCE

EPISODE 3: EXPLORATORY DATA ANALYSIS

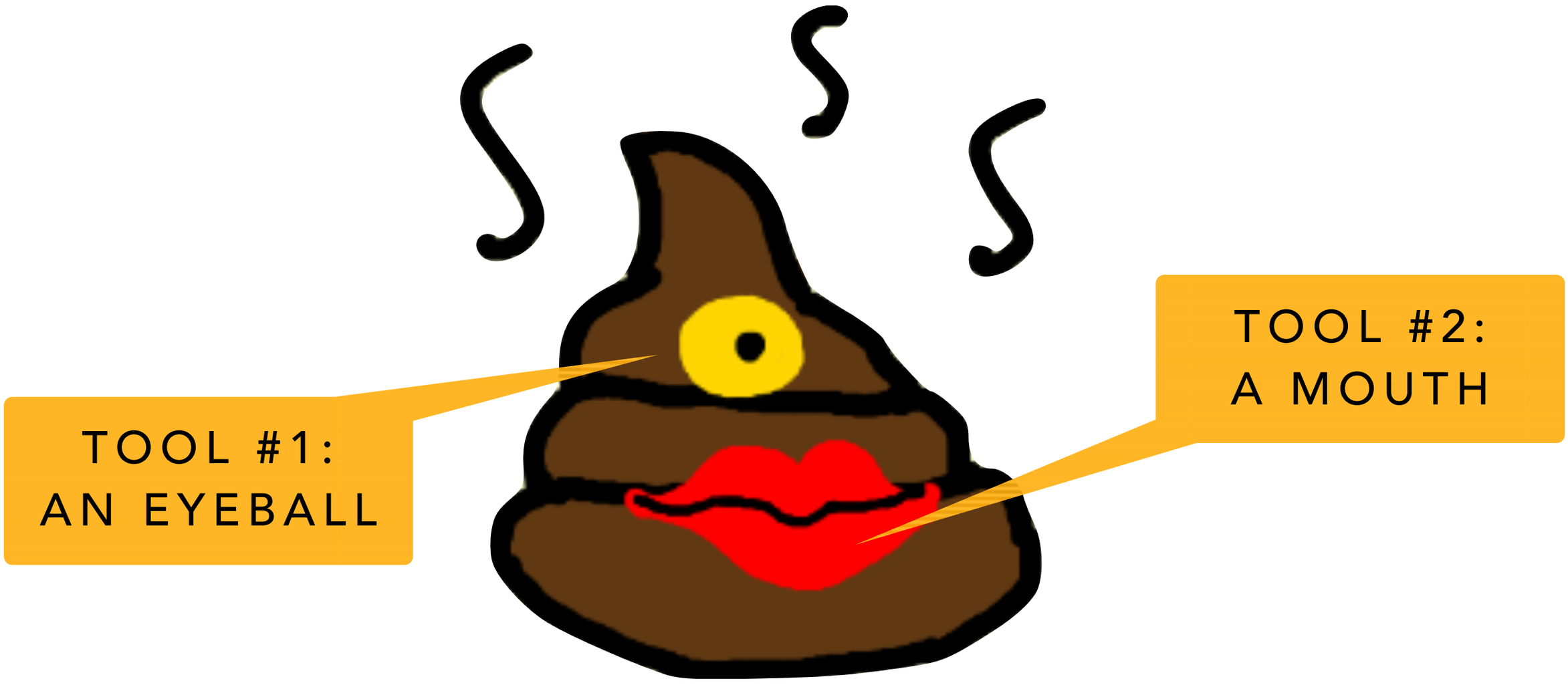
TODAY'S MENU

1. EXPLORATORY DATA ANALYSIS

2. VISUALIZATION



MEET MR DATA SCHEINTIST



EXPLORATORY ANALYSIS

- "[P]rocedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

(John Tukey, The Future of Data Analysis, July 1961)

- Tukey found that too much emphasis was put on hypothesis testing (confirmatory analysis) and that a new kind of analysis was needed where hypotheses are suggested by the data

- The development of tools like S, S-Plus, and R was motivated by Tukey's ideas

EXPLORATORY ANALYSIS

- **LOOK** AT THE DATA!
- Big problems have been caused by not looking
- If you want to be sure, go look at the **actual thing** rather than the data (check cables connections, etc.)
- Second best would be to look at the raw data
- The further you get from the source, the higher the risk that you see artefacts of the pre-processing

NON-GRAPHICAL EDA

- Again: **LOOK** AT THE DATA
- Raw data:

```
player_name,height,weight,p_id,id,player_fifa_api_id
"Abel Tamata",182.88,170,240556,1031,202153,240556,"
Abel,177.8,165,40938,1046,17880,40938,"2010-08-30 00
"Abella Perez Damia",187.96,174,37422,1051,159580,37
"Abiola Dauda",180.34,165,114503,1076,187175,114503,
"Abou Diaby",193.04,168,27277,1092,163423,27277,"201
"Aboubacar Tandia",193.04,185,181344,1119,138675,181
"Aboubakar Kamara",177.8,168,581141,1122,225541,5811
```

- Observations:
  - height in metric units [cm] but weight in imperial [lb]
  - name can be 1–3 words (at least)

NON-GRAPHICAL EDA

- Again: **LOOK** AT THE DATA
- Raw data:

```
,70,right,,_0,60,50,60,74,,74,,53,62,73,74,70,,63,,6
3,right,,_0,49,64,65,54,58,58,46,44,27,64,69,54,66,5
:00",63,72,left,medium,medium,42,52,39,64,40,72,67,4
",56,62,right,high,medium,54,37,48,62,31,56,35,39,56
3,right,medium,medium,57,53,58,69,57,59,69,68,66,64,
,right,,_0,56,34,70,70,,56,,26,71,60,69,68,,70,,72,,
",61,65,right,,_0,42,60,54,52,,51,,42,35,60,74,77,,5
,left,medium,medium,42,30,61,59,37,47,41,32,67,50,67
```

- Observations:
  - height in metric units [cm] but weight in imperial [lb]
  - name can be 1–3 words (at least)
  - missing data
  - what are these '_0'?

NON-GRAPHICAL EDA

- For table-formatted data (csv), a spreadsheet can be useful (but mind the formatting)

Player stats

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

	A	B	C	D	E	F	G	H	I
1	player_name	height	weight	p_id	id	player_fifa_api_i	player_api_id	date	overall_rating
2	Aaron Appindangoye	182.88	187	505942	1	218353	505942	2016-02-18 0:00:	67
3	Aaron Cresswell	170.18	146	155782	6	189615	155782	2016-04-21 0:00:	74
4	Aaron Doran	170.18	163	162549	39	186170	162549	2016-01-07 0:00:	65
5	Aaron Galindo	182.88	198	30572	65	140161	30572	2016-04-21 0:00:	69
6	Aaron Hughes	182.88	154	23780	88	17725	23780	2015-12-24 0:00:	70
7	Aaron Hunt	182.88	161	27316	113	158138	27316	2016-04-28 0:00:	77
8	Aaron Kuhl	172.72	146	564793	140	221280	564793	2016-04-21 0:00:	61
9	Aaron Lennon	165.1	139	30895	147	152747	30895	2015-10-16 0:00:	77
10	Aaron Lennox	190.5	181	528212	173	206592	528212	2016-02-25 0:00:	48
11	Aaron Meijers	175.26	170	101042	180	188621	101042	2015-12-03 0:00:	69
12	Aaron Mokoena	182.88	181	23889	199	47189	23889	2012-02-22 0:00:	68

GOOGLE SPREADSHEET

NON-GRAPHICAL EDA

- For table-formatted data (csv), a spreadsheet can be useful (but mind the formatting)

291	Alberto Fontana	185.42	161	39425	484
292	Alberto Frison,18	190.5	187	30143	484
293	Alberto Garcia	182.88	170	102394	487
294	Alberto Gilardino	182.88	174	30881	488
295	Alberto Giuliatto	180.34	170	42460	492
296	Alberto Grassi	182.88	165	575364	493
297	Alberto Guitian	182.88	163	543020	494
298	Alberto Lopo	185.42	179	37451	496
299	Alberto Luque,21	182.88	176	32763	498
300	Alberto Maria Fontana	187.96	183	39743	499
301	Alberto Moreno	170.18	143	314605	500

NON-GRAPHICAL EDA

- For table-formatted data (csv), a spreadsheet can be useful (but mind the formatting)

63	68	right	medium	high	36
62	66	right	le	ean	61
71	76	right	medium	medium	70
73	73	right	low	high	57
70	70	right	medium	medium	63
73	76	left	medium	low	47
69	71	right	medium	medium	68
61	70	right	medium	medium	58
68	71	right	high	medium	63
73	75	right	None	o	73
71	71	right	high	medium	82

EXPLORATORY DATA ANALYSIS (EDA)

- There are various classes of EDA:
  - **non-graphical** vs **graphical**
  - **univariate** vs **bivariate** vs **multivariate**
  - etc

- You should use them all

- The basic goals are:

1. **remove or correct erroneous data**
2. **formulate initial hypotheses**
3. **choose suitable analysis methods**

WHY EDA

- Any suspicious data need to be investigated and corrected/ removed

- This can be quite tricky: is it an **outlier** or a key observation?

- After EDA, it may be easier to formulate hypotheses, and to make an informed choice of the analysis approach

- E.g.:
  - choose between linear vs non-linear methods
  - decide whether clustering would be helpful

SUMMARY STATISTICS

- Simple summary statistics:
  - continuous: average, median, min, max
  - categorical: set of values, mode

R

```
> D <- read.csv("player_stats.csv")
> summary(D)
  player_name                height          weight                p
 Danilo   :     7   Min.   :157.5   Min.   :117.0   Min.
 Paulinho:     6   1st Qu.:177.8   1st Qu.:159.0   1st Qu
 Ricardo :     5   Median :182.9   Median :168.0   Median
 Adriano :     4   Mean   :181.9   Mean   :168.4   Mean
 Douglas :     4   3rd Qu.:185.4   3rd Qu.:179.0   3rd Qu
 Felipe  :     4   Max.   :208.3   Max.   :243.0   Max.
 (Other) :11034
       id          player_fifa_api_id player_api_id
 Min.   :     1   Min.   :     2     Min.   :  2625
 1st Qu.: 46173   1st Qu.:151895     1st Qu.: 35558
 Median : 92080   Median :184705     Median : 96622
 Mean   : 92278   Mean   :165686     Mean   :156573
 3rd Qu.:138874   3rd Qu.:203884     3rd Qu.:212442
 Max.   :183969   Max.   :234141     Max.   :750584
```

SUMMARY STATISTICS

- Simple summary statistics:
  - continuous variables: *average, median, min, max*
  - categorical variables: *set of values, mode*

```
attacking_work_rate defensive_work_rate     crossing
medium  :6967       medium  :7311        Min.   : 6.00
high    :2375       high    :1555        1st Qu.:44.00
low     : 565       low     :1041        Median :58.00
          544       _0      : 540        Mean   :53.98
N        495       o       : 278        3rd Qu.:67.00
norm     :  66       ormal   :  66        Max.   :92.00
(Other) :  52       (Other) : 273        NA's   :4
```

WHY TABLES ARE MUCH BETTER THAN GRAPHS (APRIL 1ST, 2009)

- "And I recommend using Excel, which has some really nice defaults as well as options such as those 3-D colored bar charts. "

(Andrew Gelman, "Why Tables are Really Much Better than Graphs", *Journal of Computational and Graphical Statistics*, Volume 20, Number 1, Pages 3–7

Method	*Saturated fat* (≤ 30 *g/day*; ≥ 30 *g/day*)	*Total energy intake* (*1,000 kcal/day*)	*Alcohol* (*20 g/day*; ≥ 20 *g/day*)	*Age* (*10 years*)
UC	.94 (.75–1.17)	.95 (.78–1.16)	1.35 (1.05–1.74)	1.60 (1.43–1.87)
RC	.78 (.33–1.80)	.80 (.27–2.38)	1.57 (1.09–2.28)	1.54 (1.42–1.67)
ML-WICI	.65 (.18–2.44)	.89 (.26–3.05)	1.55 (1.06–2.27)	1.53 (1.33–1.76)
ML-WGCI	.65 (.22–1.99)	.89 (.26–3.01)	1.55 (1.05–2.27)	1.53 (1.33–1.77)
ML-PLCI	.65 (0–2.13)	.89 (.19–3.61)	1.55 (1.07–2.60)	1.53 (1.25–1.75)
ML-RbCI	.65 (.13–3.22)	.89 (.24–3.30)	1.55 (1.05–2.28)	1.53 (1.34–1.75)
GEE a*-RBCI	.71 (.21–2.38)	.88 (.25–3.15)	1.53 (1.05–2.25)	1.53 (1.38–1.75)
GEE b-RbCI	.69 (.15–3.10)	.93 (.20–4.42)	1.52 (1.01–2.30)	1.55 (1.37–1.76)
GEE c-RbCI	.69 (.19–2.44)	.94 (.24–3.71)	1.52 (1.03–2.23)	1.55 (1.38–1.75)

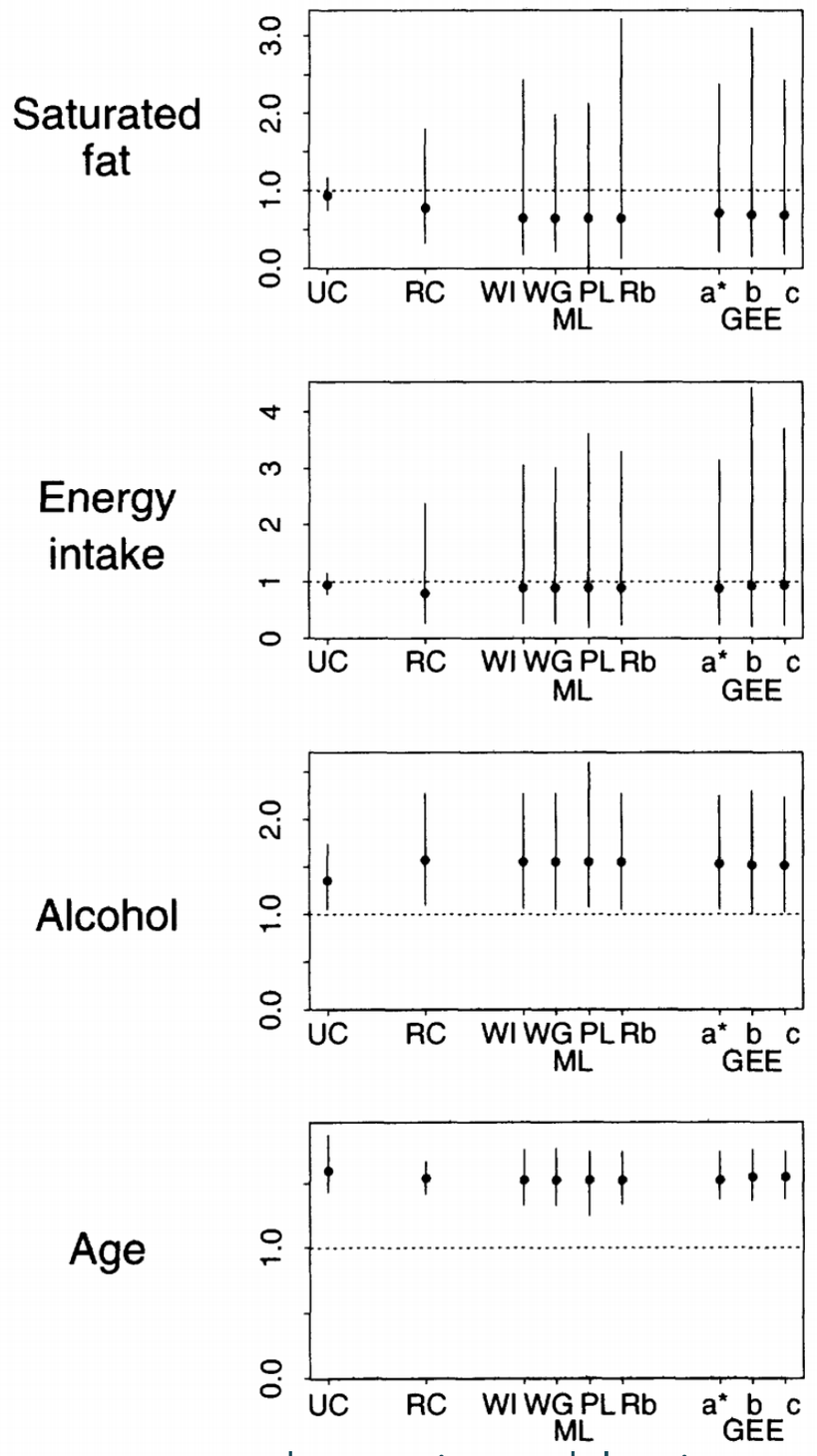
Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. "Let's practice what we preach: turning tables into graphs." The American Statistician 56.2 (2002): 121-130.

WHY TABLES ARE MUCH BETTER THAN GRAPHS (APRIL 1ST, 2009)

- "And I recommend using Excel, which has some re as options such as those 3-D colored bar charts. "

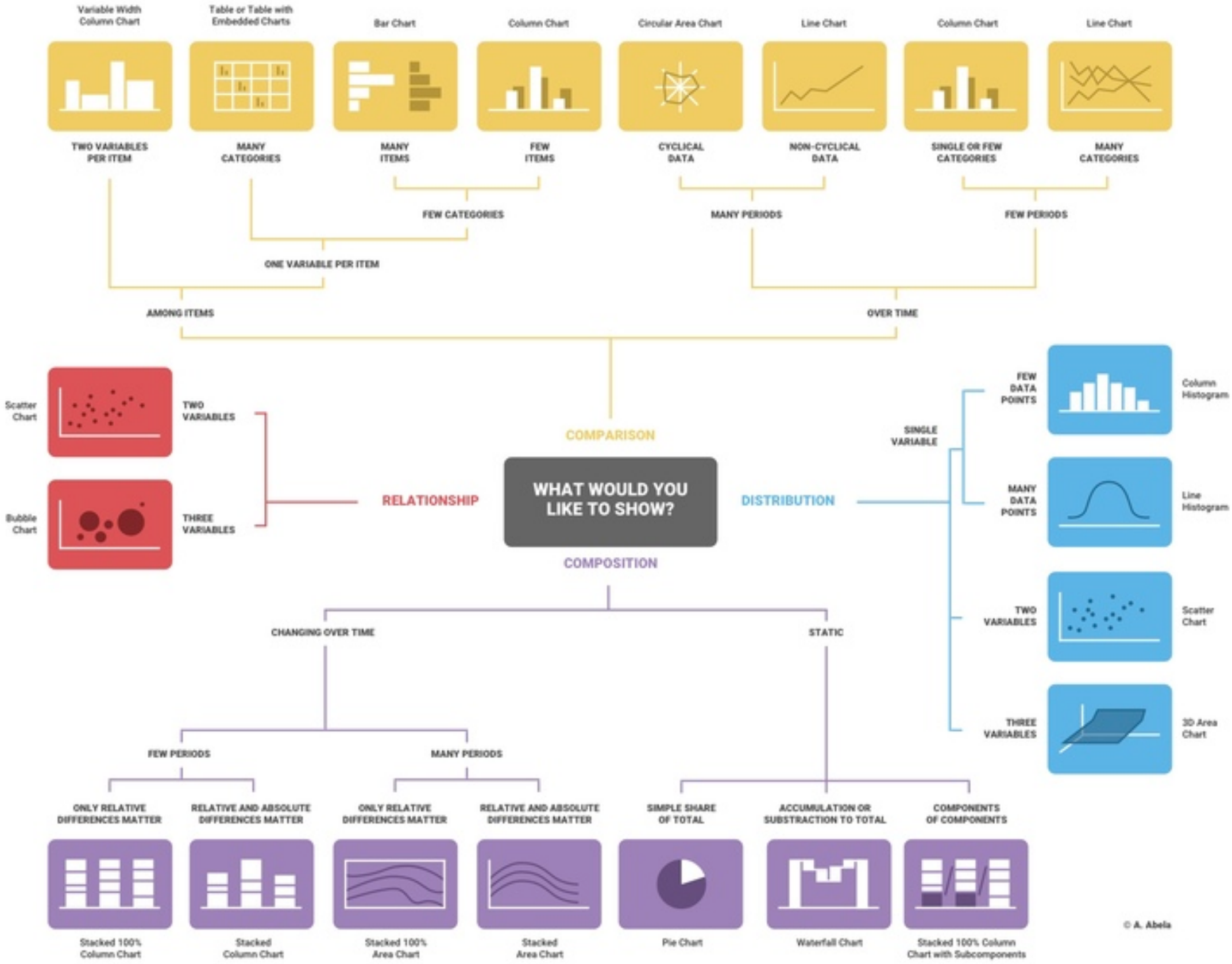
(Andrew Gelman, "Why Tables are Really Much Better than Graph *and Graphical Statistics*, Volume 20, Number 1, Pages 3–7

Method	*Saturated fat* (≤ 30 *g/day*; ≥ 30 *g/day*)	*Total energy intake* (*1,000 kcal/day*)	(2
UC	.94 (.75–1.17)	.95 (.78–1.16)	1.
RC	.78 (.33–1.80)	.80 (.27–2.38)	1.
ML-WICI	.65 (.18–2.44)	.89 (.26–3.05)	1.
ML-WGCI	.65 (.22–1.99)	.89 (.26–3.01)	1.
ML-PLCI	.65 (0–2.13)	.89 (.19–3.61)	1.
ML-RbCI	.65 (.13–3.22)	.89 (.24–3.30)	1.
GEE a*-RBCI	.71 (.21–2.38)	.88 (.25–3.15)	1.
GEE b-RbCI	.69 (.15–3.10)	.93 (.20–4.42)	1.
GEE c-RbCI	.69 (.19–2.44)	.94 (.24–3.71)	1.



Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. "Let's practice what we preach: turning tables into graphs." The American Statistician 56.2 (2002): 121-130.

CHOOSING THE RIGHT KIND OF CHART

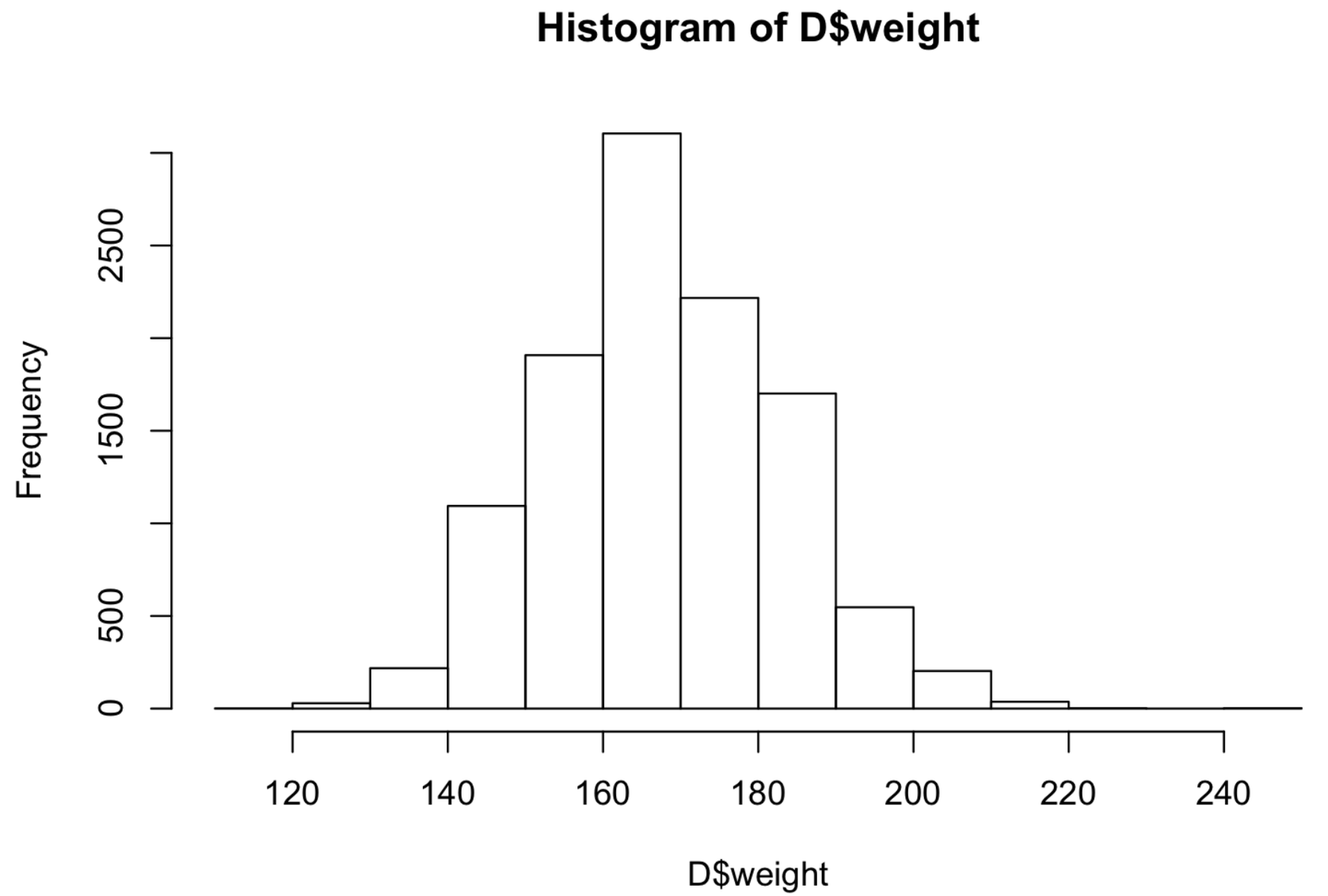


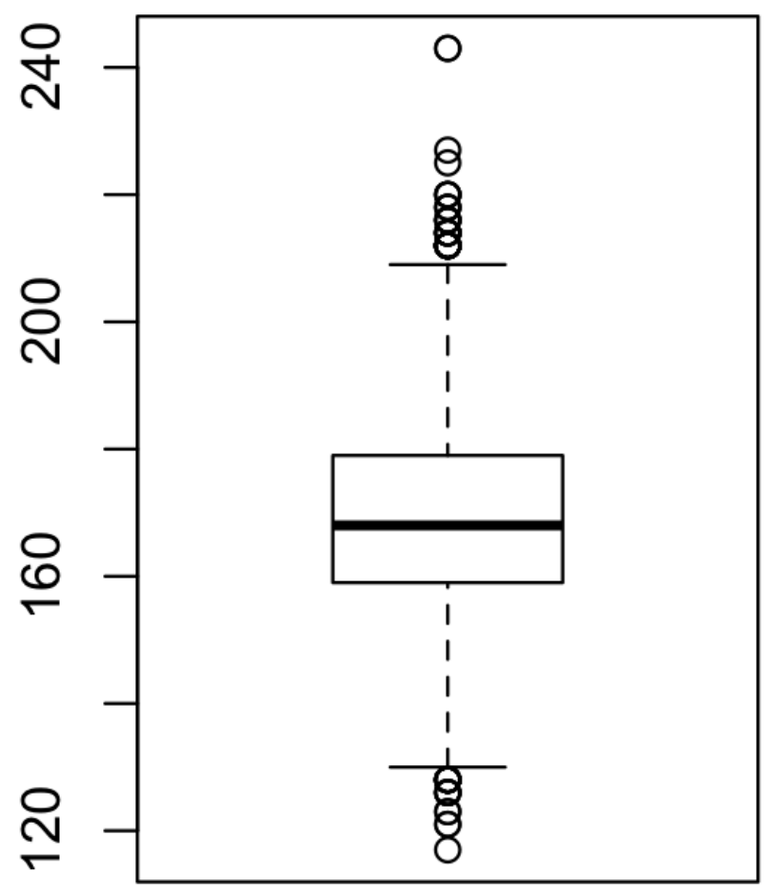
infogram.com: *How to Choose the Right Chart for Your Data*

GRAPHICAL EDA

- Distribution of a single quantity: histrogram, boxplot

```
> D <- read.csv("player_stats.csv")
> hist(D$weight)
> boxplot(D$weight)
```

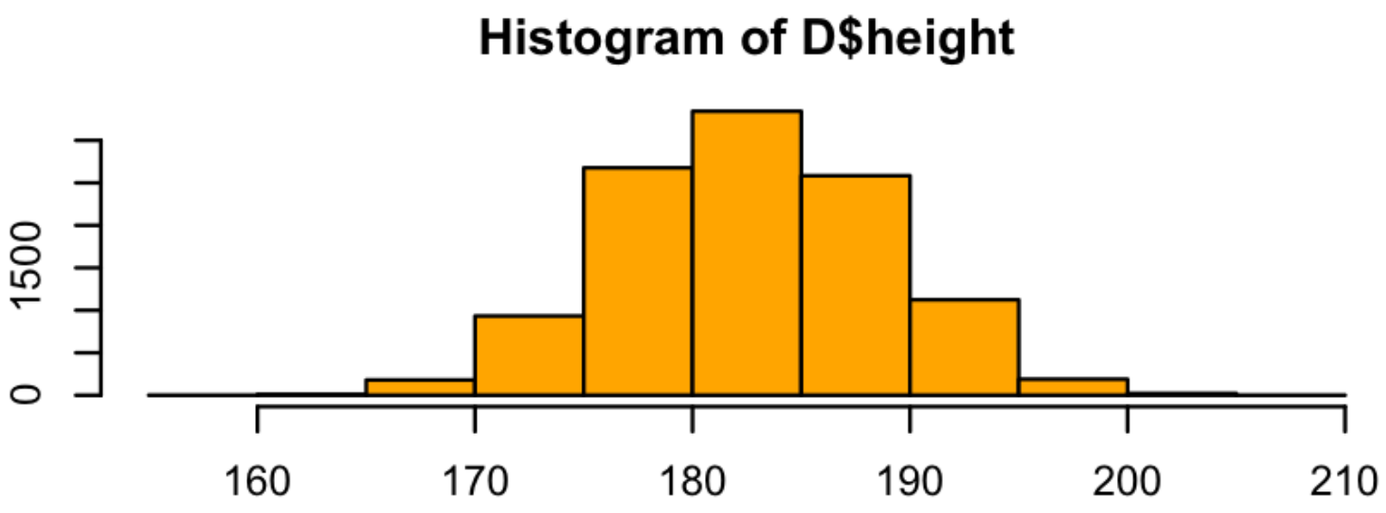




ON HISTOGRAMS

- Choosing the bin width may change how things look

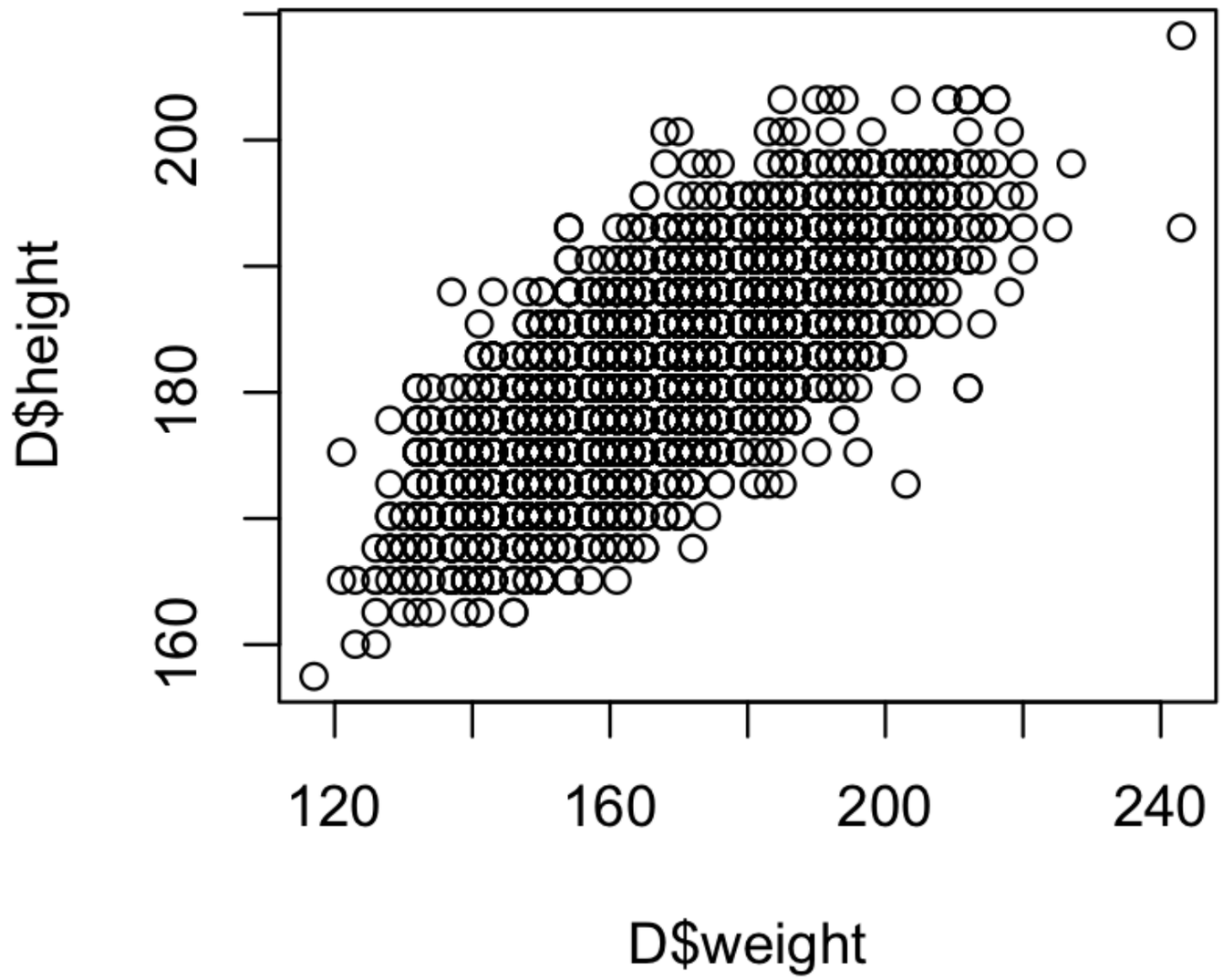
```
> par(mfrow=c(3,1), mar=c(2.5,2.5,2.5,0.5))
> hist(D$height,breaks=10,col='orange')
> hist(D$height,breaks=20,col='orange')
```



GRAPHICAL EDA

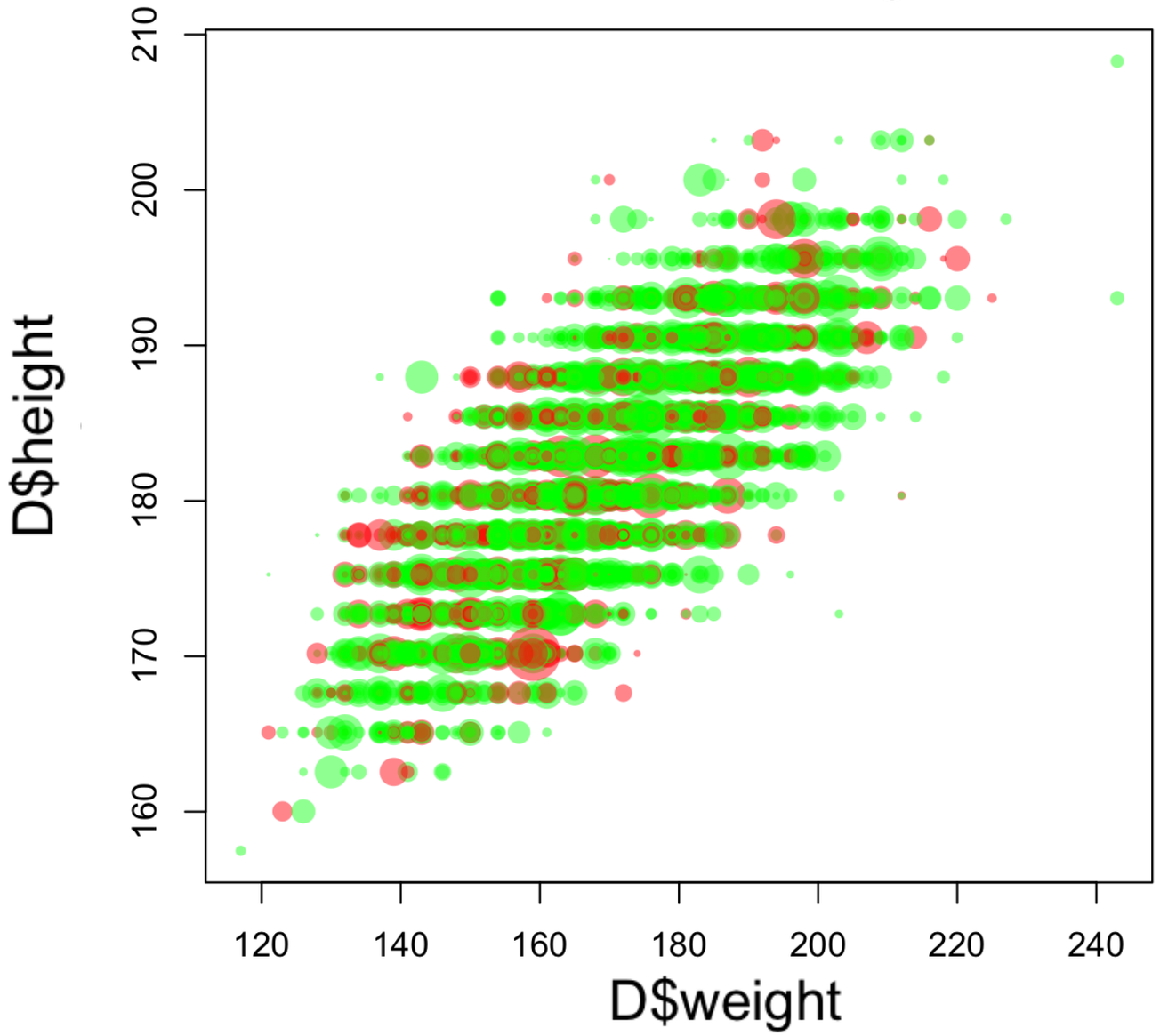
- relationship between two variables: scatter plot

```
> D <- read.csv("player_stats.csv")
> plot(D$weight, D$height)
```



GRAPHICAL EDA

- relationship between two variables: scatter plot
- more variables can be included using color and size



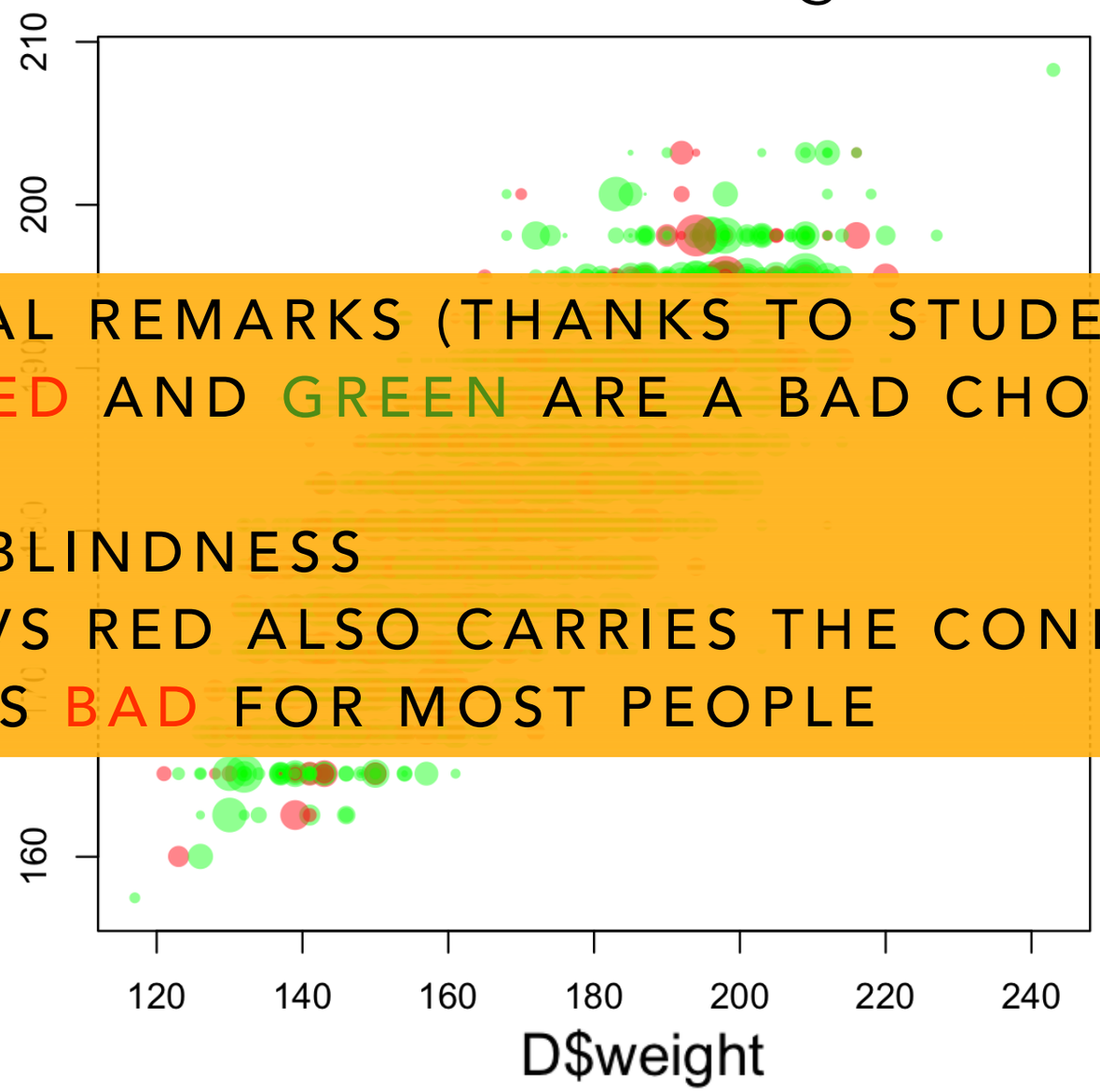
red = left footed

green = right footed

size = overall rating

GRAPHICAL EDA

- relationship between two variables: scatter plot
- more variables can be included using color and size



red = left footed

green = right footed

rating

ADDITIONAL REMARKS (THANKS TO STUDENTS!)
ON WHY RED AND GREEN ARE A BAD CHOICE

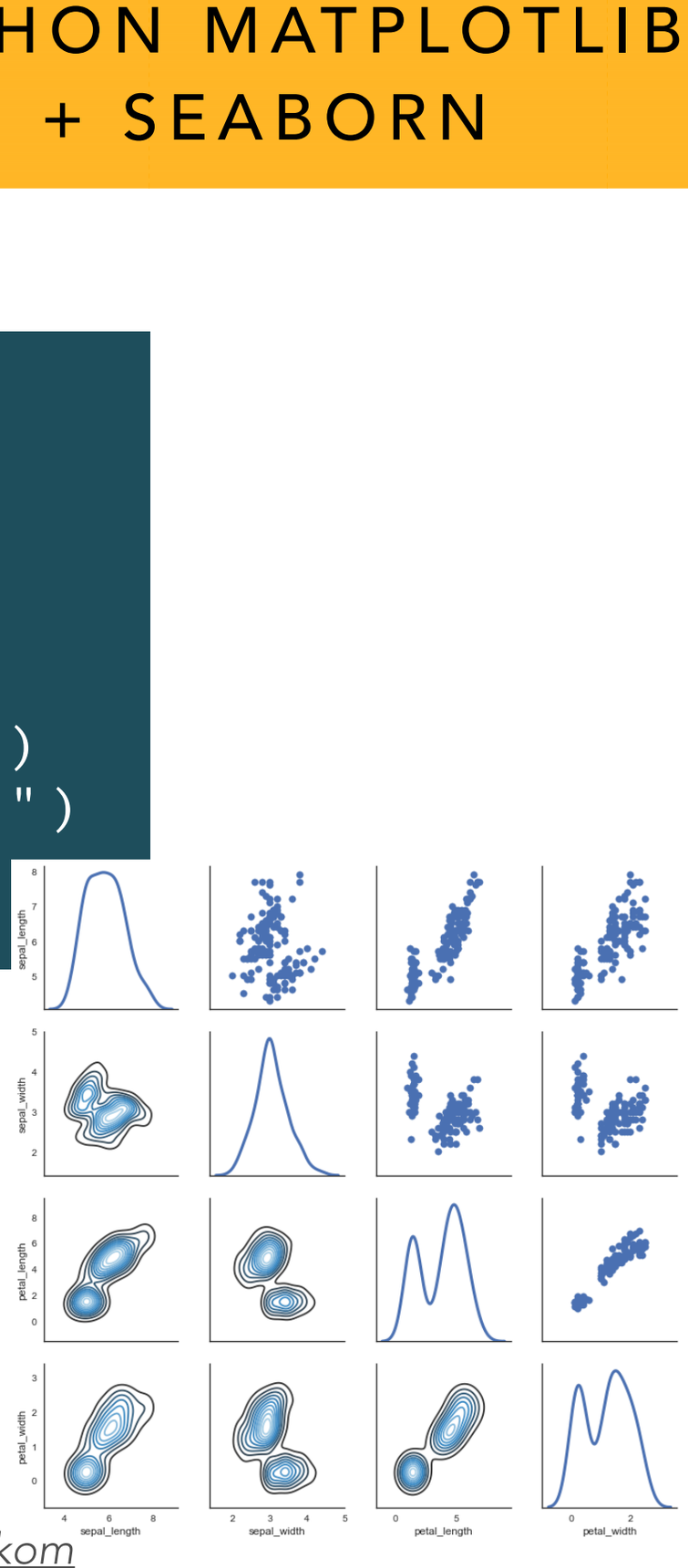
1. COLOR BLINDNESS
2. GREEN VS RED ALSO CARRIES THE CONNOTATION
   GOOD VS BAD FOR MOST PEOPLE

GRAPHICAL EDA

PYTHON MATPLOTLIB + SEABORN

- A majority of graphical EDA is 1D or 2D

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(style="white")

df = sns.load_dataset("iris")

g = sns.PairGrid(df, diag_sharey=False)
g.map_lower(sns.kdeplot, cmap="Blues_d")
g.map_upper(plt.scatter)
g.map_diag(sns.kdeplot, lw=3)
```



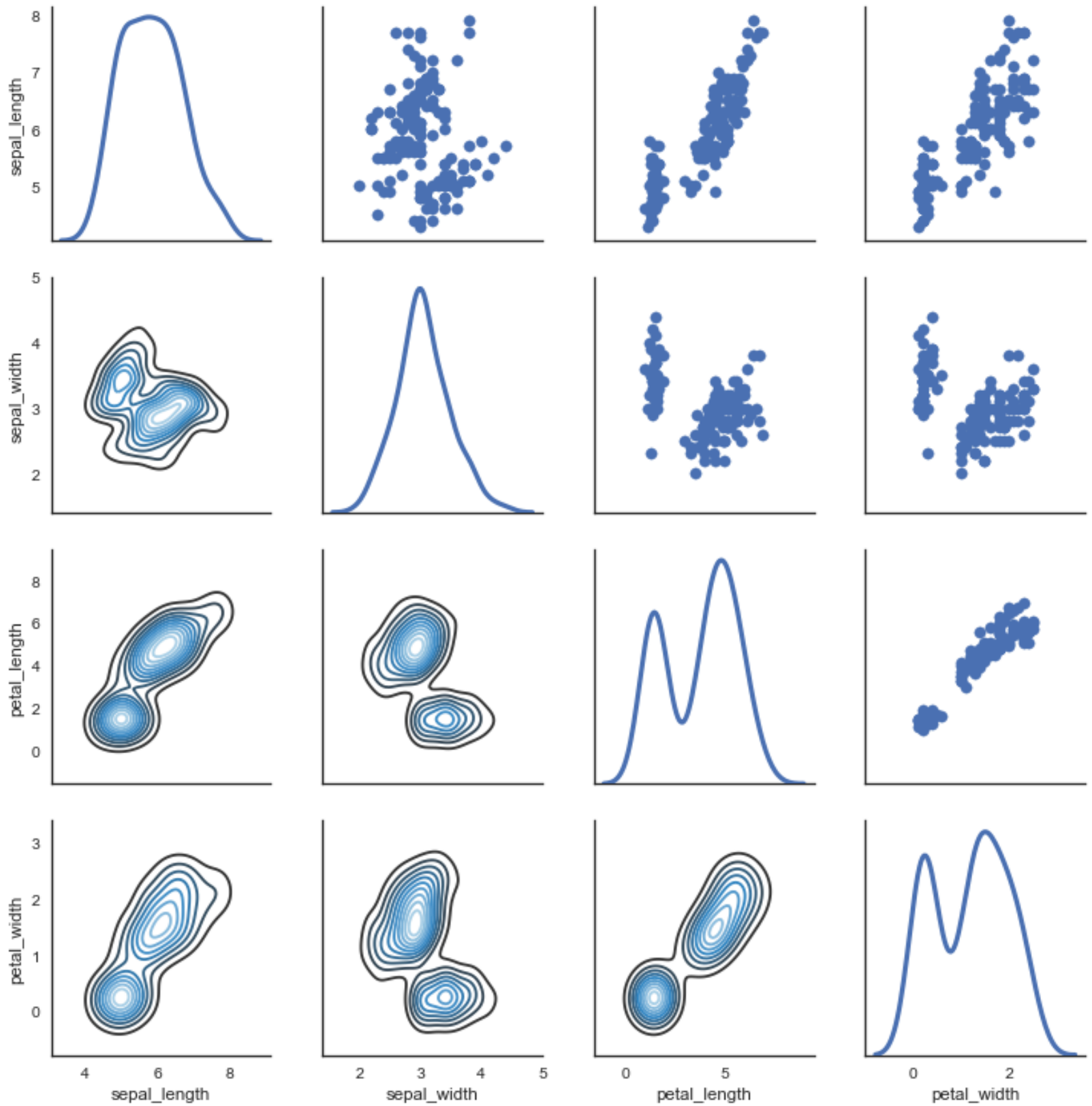
examples on this page from: "seaborn: statistical data visualization", Michael Waskom

GRAPHICAL EDA

PYTHON MATPLOTLIB

- A majority of graphic

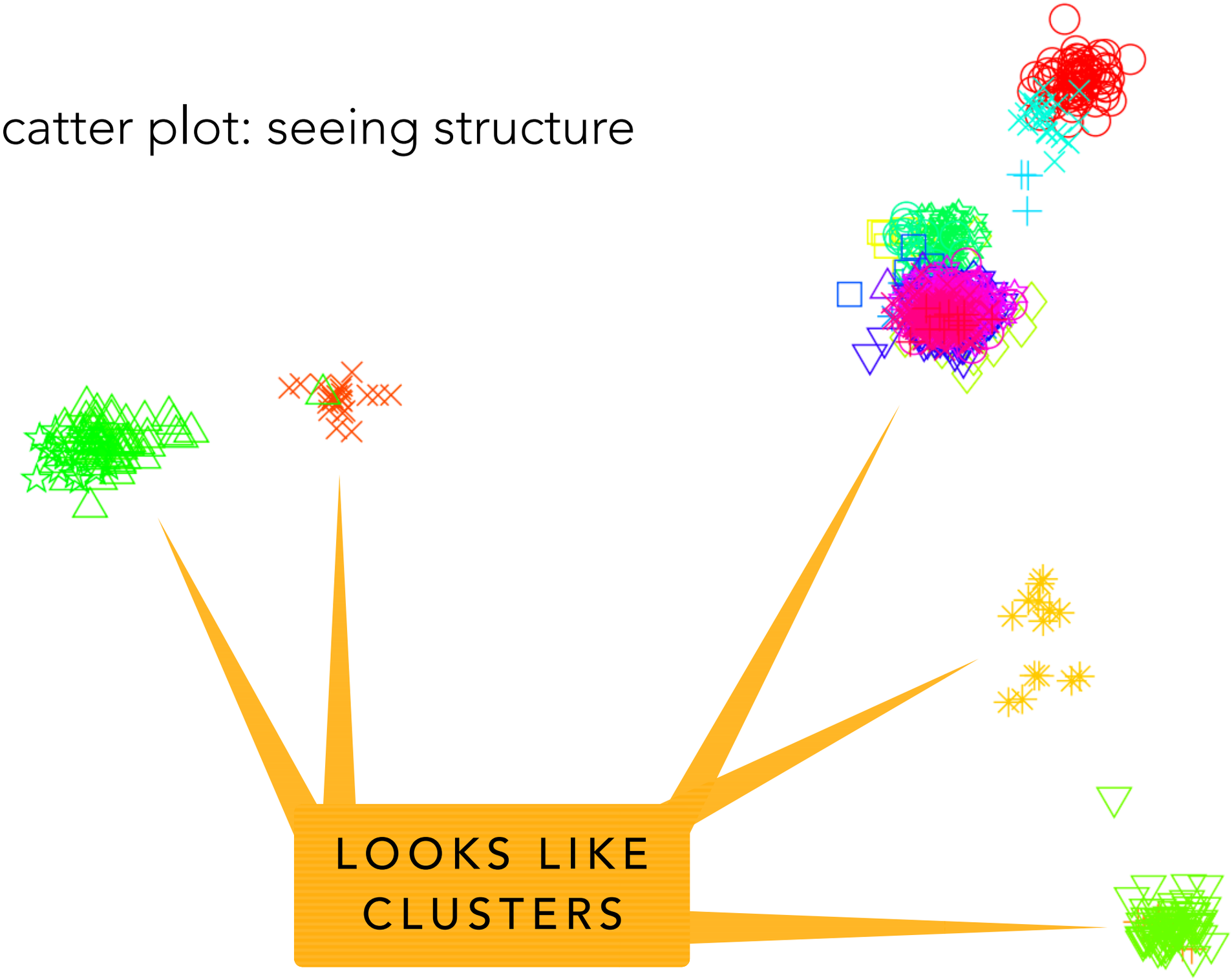
```
import seaborn as sns
import matplotlib.pyp
sns.set(style="white"

df = sns.load_dataset

g = sns.PairGrid(df,
g.map_lower(sns.kdepl
g.map_upper(plt.scatt
g.map_diag(sns.kdeplo
```



examples on this page from: "seaborn: statis

GRAPHICAL EDA

- Scatter plot: seeing structure



ABOUT STATISTICAL GRAPHICS

- Visualization techniques and even the details can be crucial in data analysis and communication of the findings

- Darrell Huff, *How to Lie with Statistics*, 1954

- Edward R Tufte, *The Visual Display of Quantitative Information*, 1983

TUFTE, 1983

- Principles of graphical **integrity:**
    1. Representation (length, area, ...) should be directly proportional to the number
    2. Clear labeling on the graphic itself – avoid legends
    3. Show data variation, not design variation
    4. Apply inflation-adjustment for money
    5. The number of information-carrying dimensions depicted should not exceed the number of dimensions in the data.

- Principles of graphical **excellence:**
    1. **Fundamental principle: Above all else show the data.**
    2. Maximize the data-ink ratio.
    3. Erase non-data ink.
    4. Erase redundant data ink.
    5. Revise and edit.

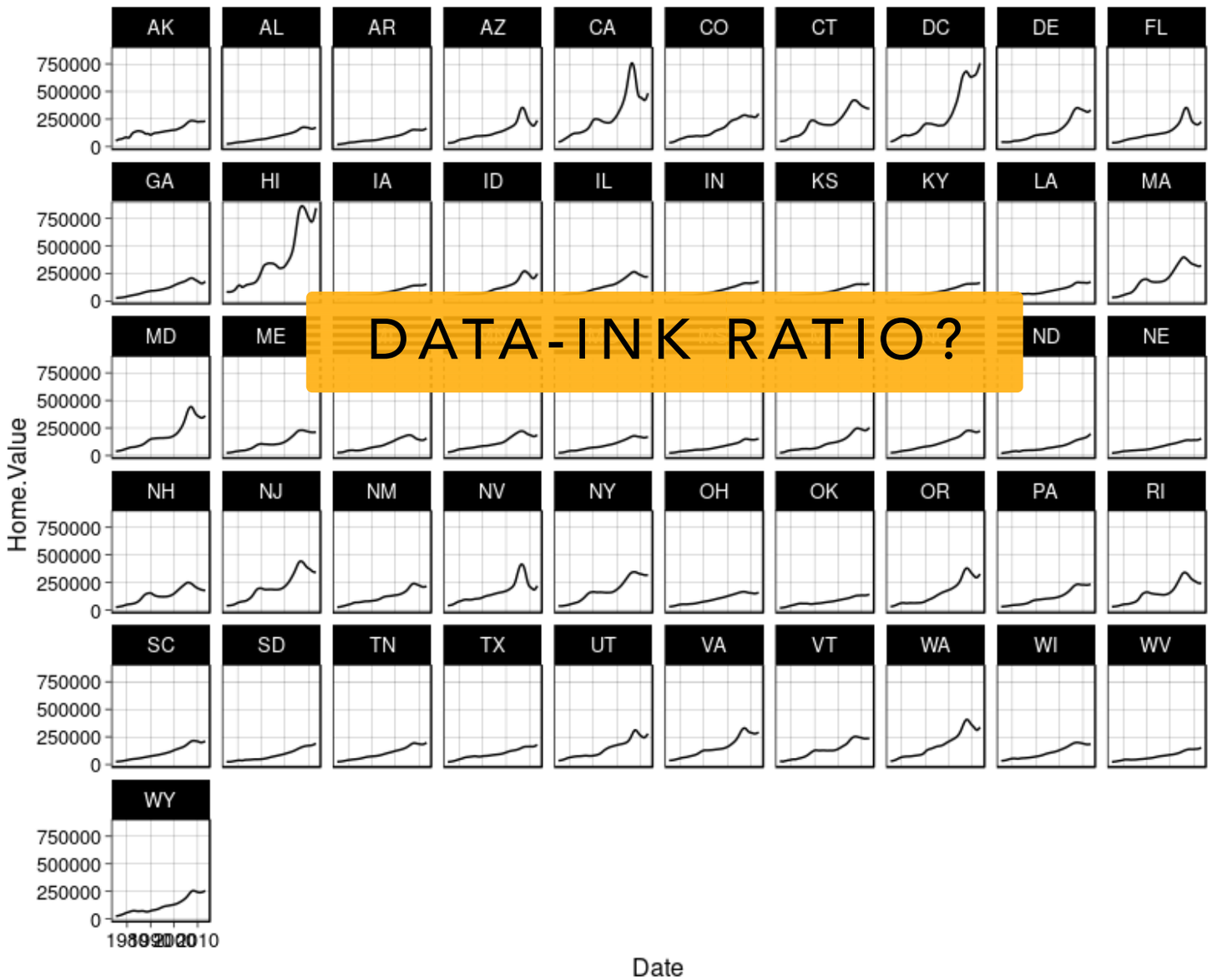
TUFTE, 1983

- "data–ink ratio = $\frac{\text{data-ink}}{\text{total ink used to print the graphic}}$

= proportion of a graphic's ink devoted to the non-redundant display of data information "

THE GOOD, THE BAD, AND THE UGLY #1

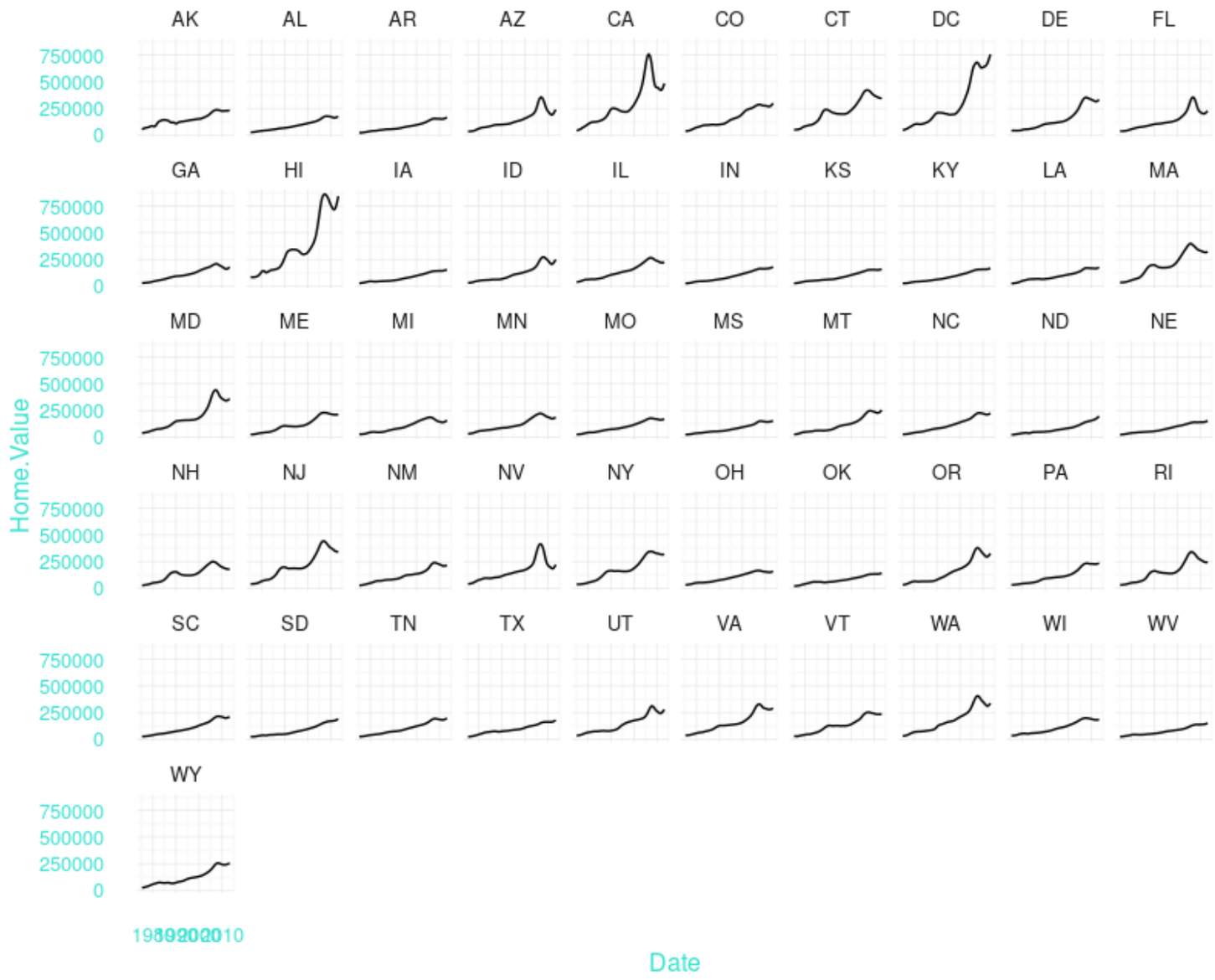
- Apparently "cosmetic" things such as line weight and decorations (borders, labels, etc) can make all the difference



examples on this page from: Introduction to R Graphics with ggplot2, Harvard University

THE GOOD, THE BAD, AND THE UGLY #1

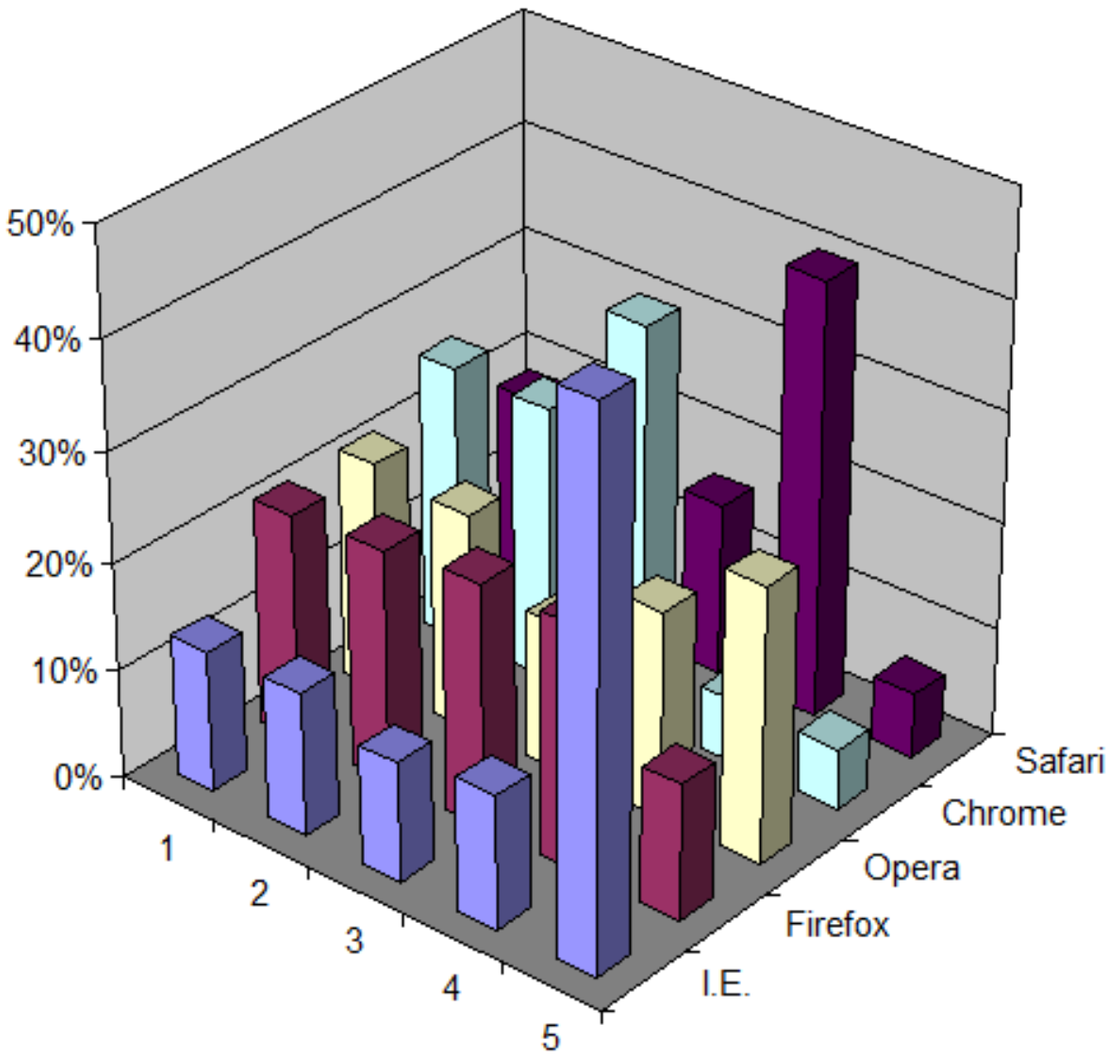
- Apparently "cosmetic" things such as line weight and decorations (borders, labels, etc) can make all the difference

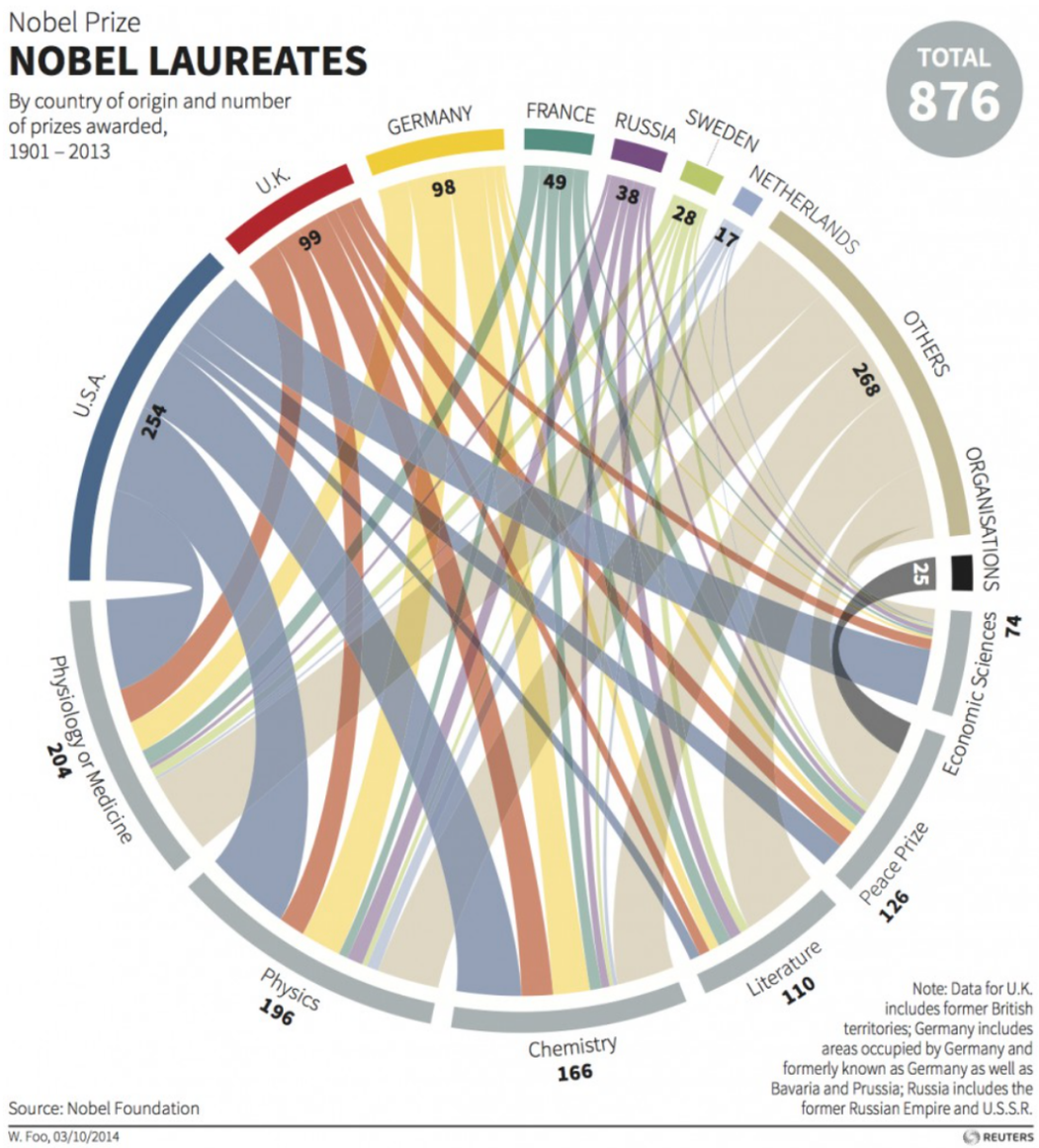


examples on this page from: Introduction to R Graphics with ggplot2, Harvard University

THE GOOD, THE BAD, AND THE UGLY #2 & #3

- Even though you can produce a really cool visualization, it's not necessarily the best way to show the data or convey the idea

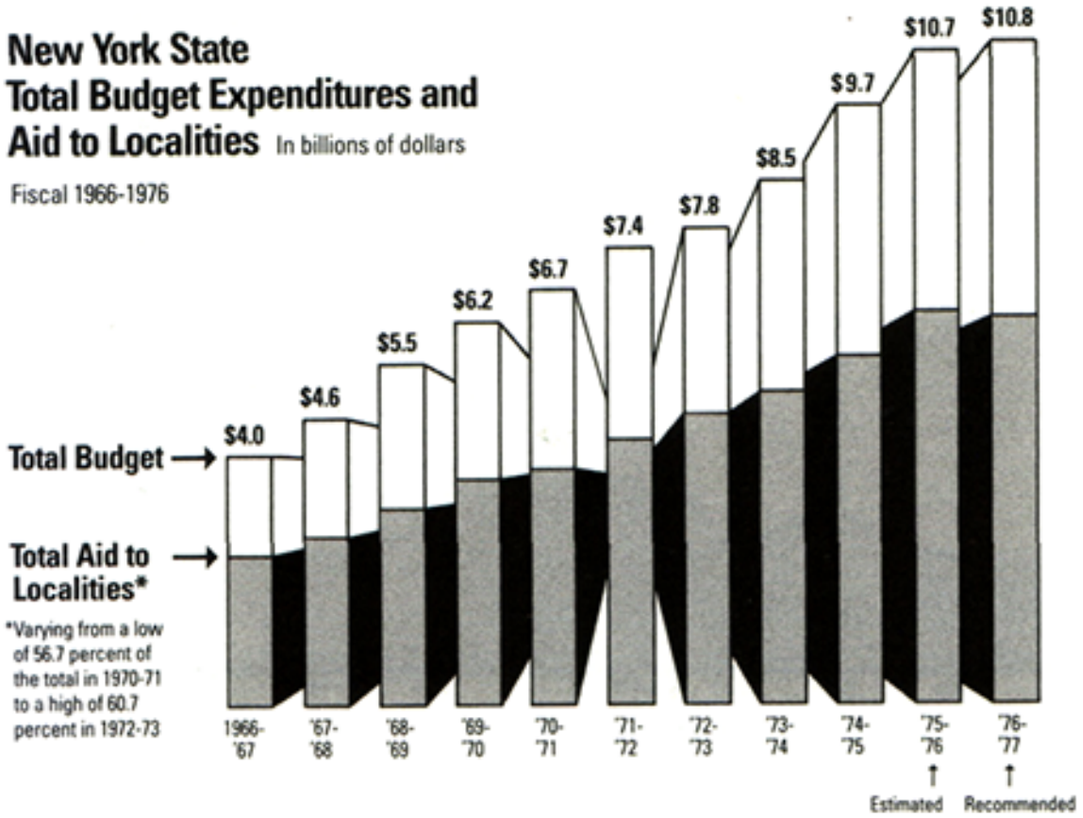


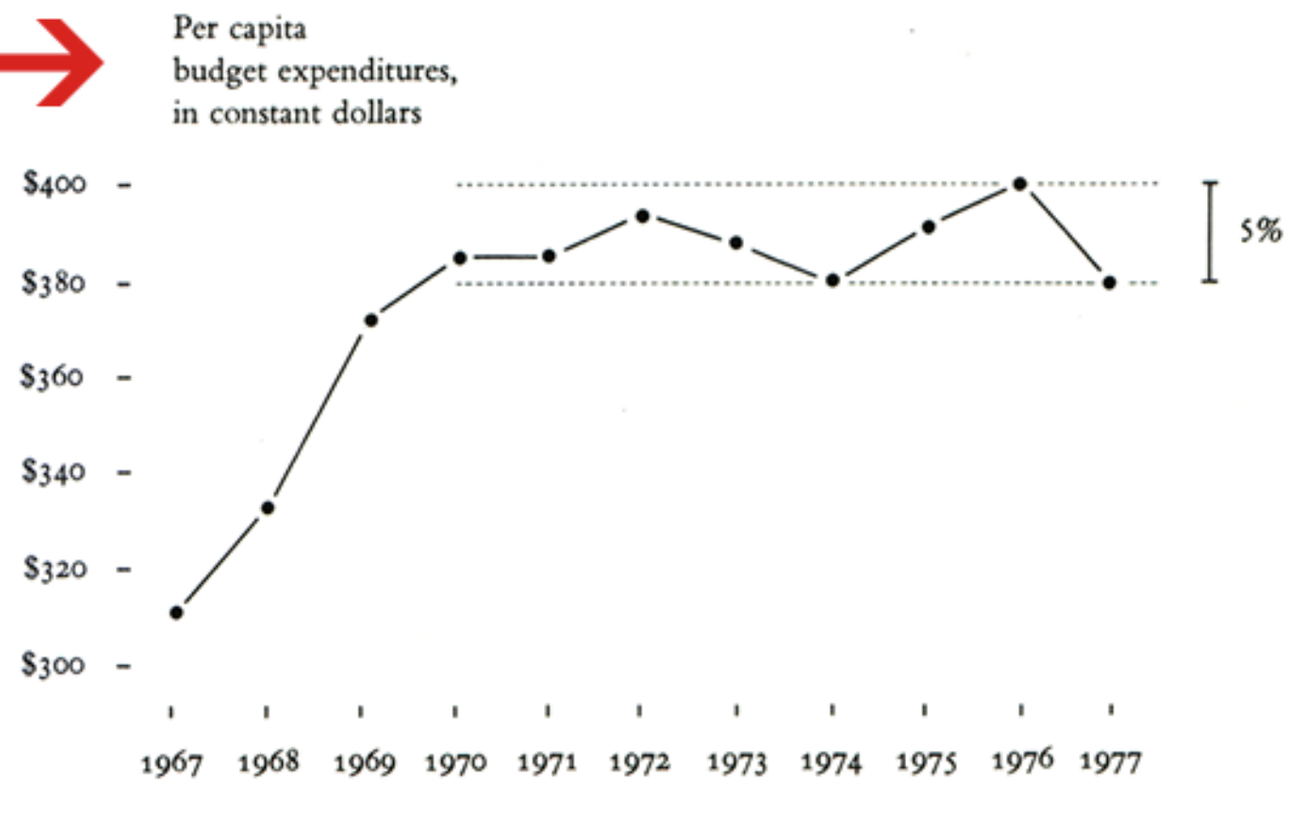


examples on this page from: English 419: Multimedia writing

THE GOOD, THE BAD, AND THE UGLY #4

- 4. Apply inflation-adjustment for money





DATA-INK RATIO?

from: (Tufte, 1983)

THE GOOD, THE BAD, AND THE UGLY #5

5. " Representation (length, area, ...) should be directly proportional to the number " (Tufte, 1983)

- If you map a one-dimensional feature (such as money, weight) to 2D, the area (not the diameter) should be directly proportional to the feature.

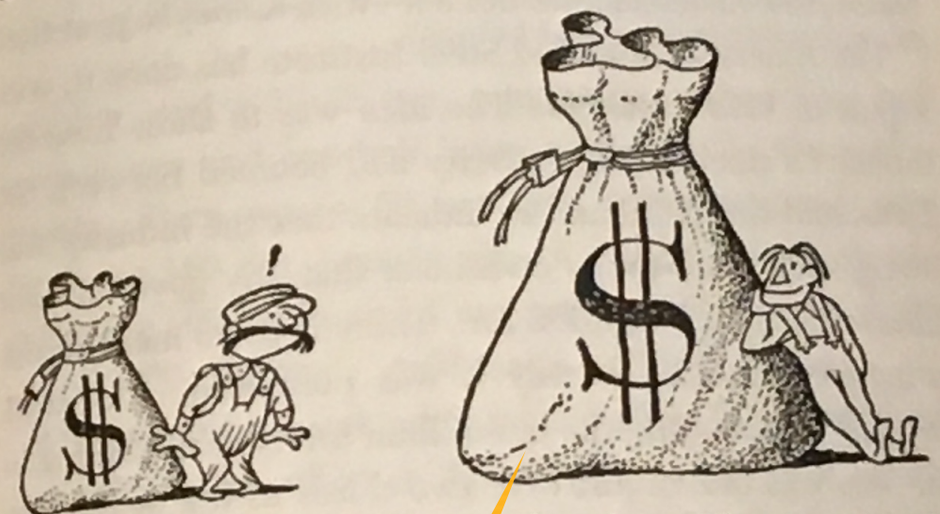
There it is, with dollars-per-week indicated up the left side. It is a clear and honest picture. Twice as much money is twice as big on the chart and looks it.

The chart lacks that eye-appeal though, doesn't it? I can easily supply that by using something that looks more like money than a bar does: moneybags. One moneybag

for the unfortunate Rotundian's pittance, two for the American's wage. Or three for the Rotundian, six for the

2X

to represent the [illegible] proportion, isn't it?



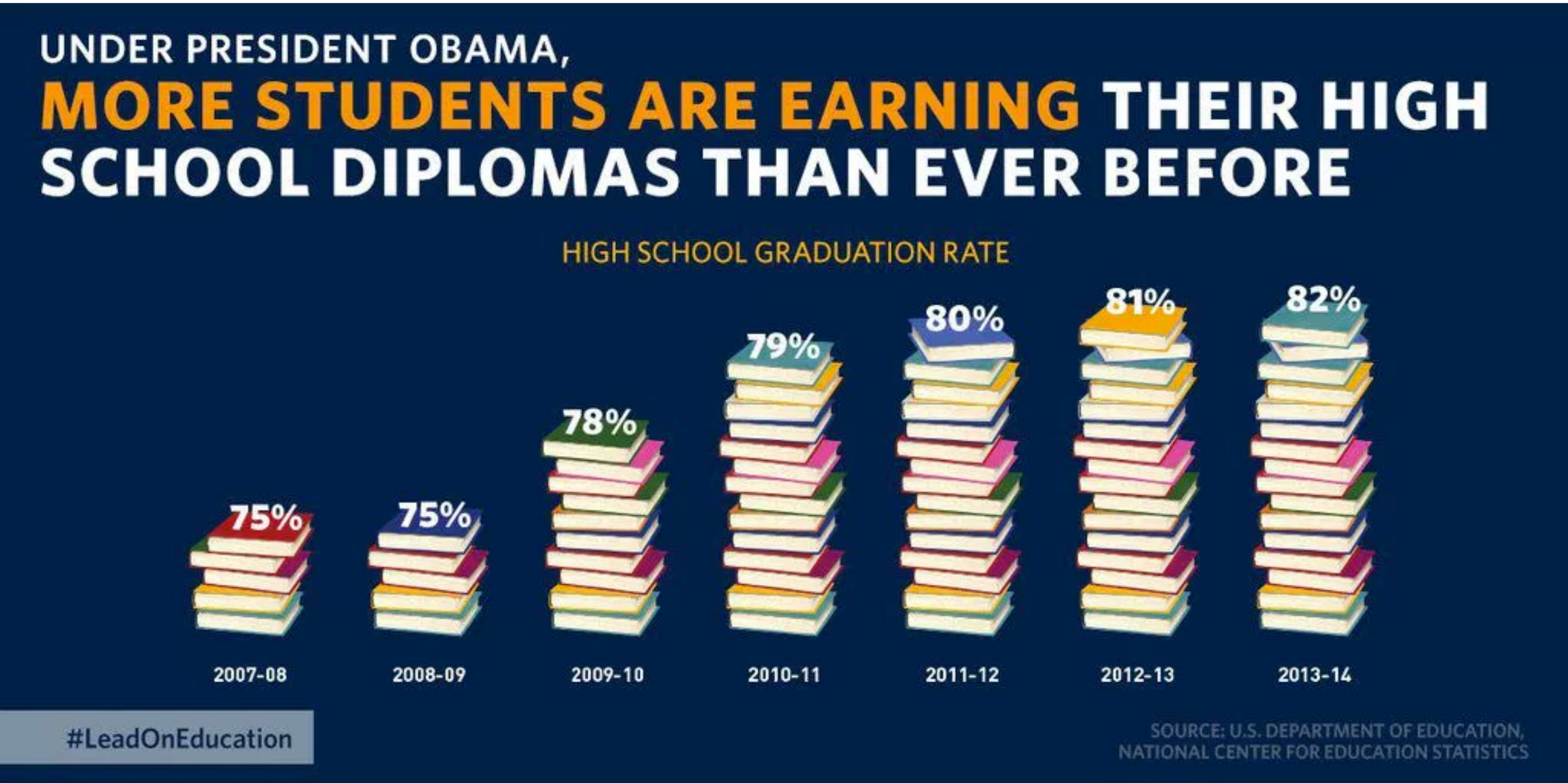
Now *that* gives the impression I'm after. The American's wage now dwarfs the foreigner's.

= 4X (OR 8X)?

from: (Darrell Huff, 1954)

THE GOOD, THE BAD, AND THE UGLY #6

- 16 books is *much* better than five books, right?

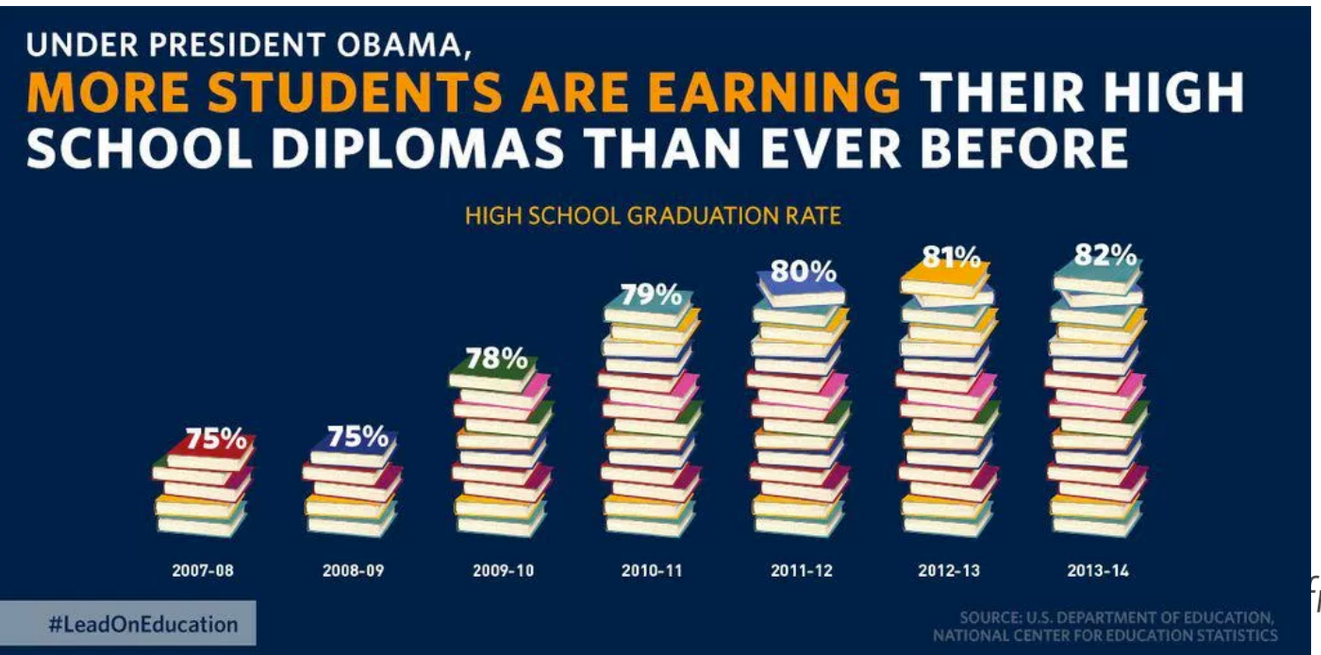


THE GOOD, THE BAD, AND THE UGLY #6

- 16 books is *much* better than five books, right?

1. Representation (length, area, ...) should be directly proportional to the number

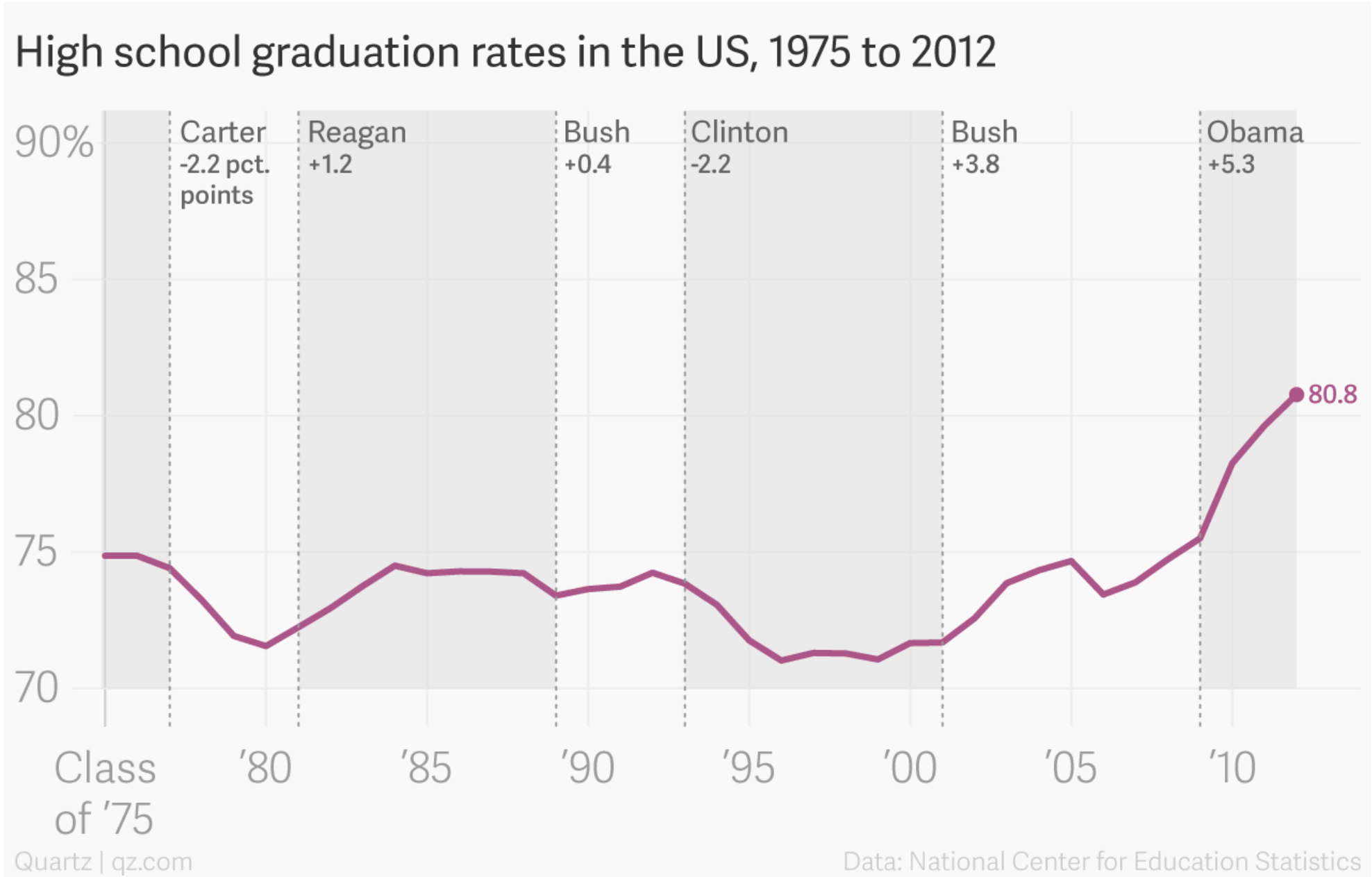
- Therefore, column (bar) charts should always start at zero!



from: "The most misleading charts of 2015, fixed", qz.com

THE GOOD, THE BAD, AND THE UGLY #6

- 16 books is *much* better than five books, right?



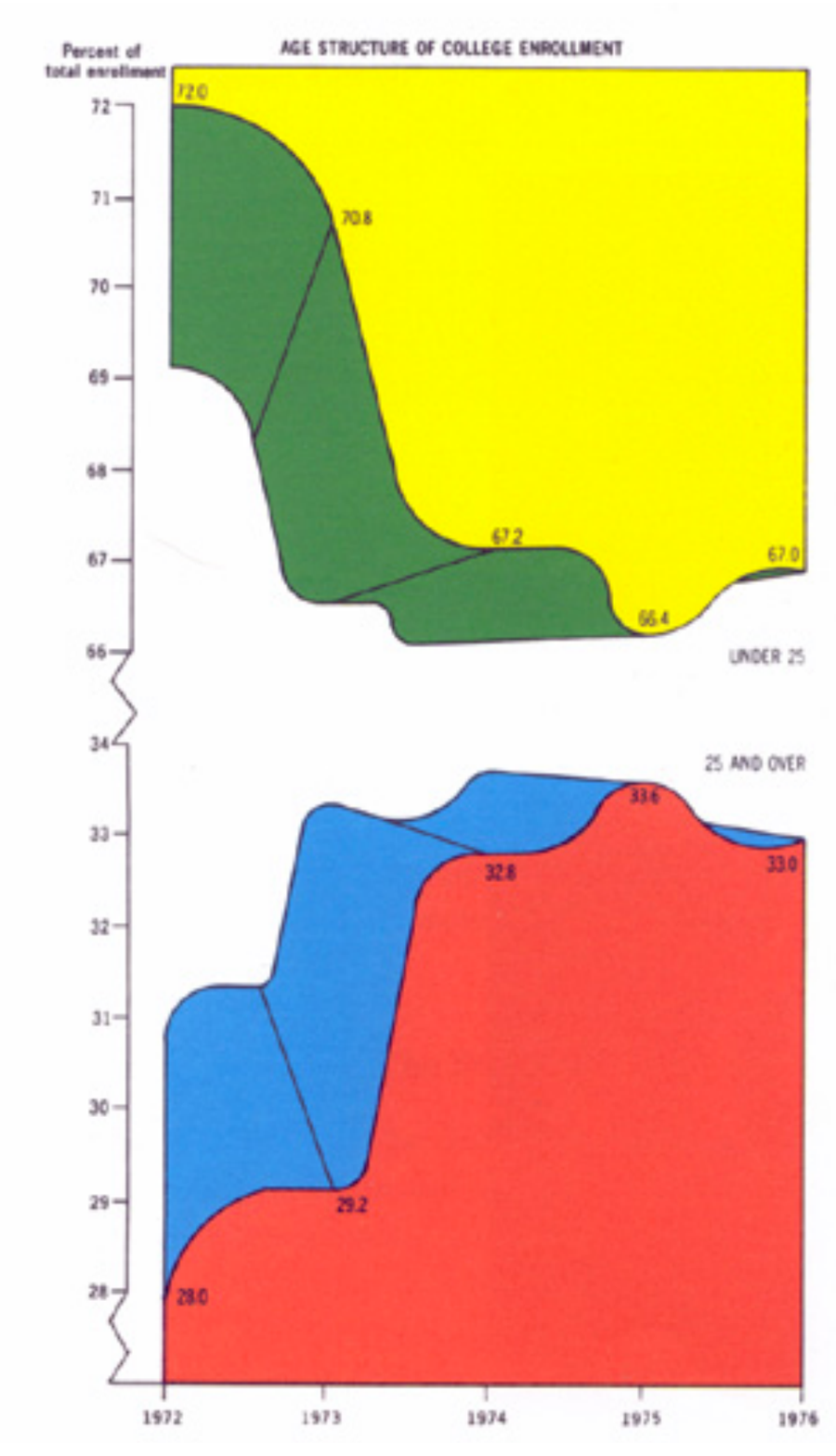
examples on this page from: "The most misleading charts of 2015, fixed", qz.com

THE GOOD, THE BAD, AND THE UGLY #7

TUFTE, 1983:

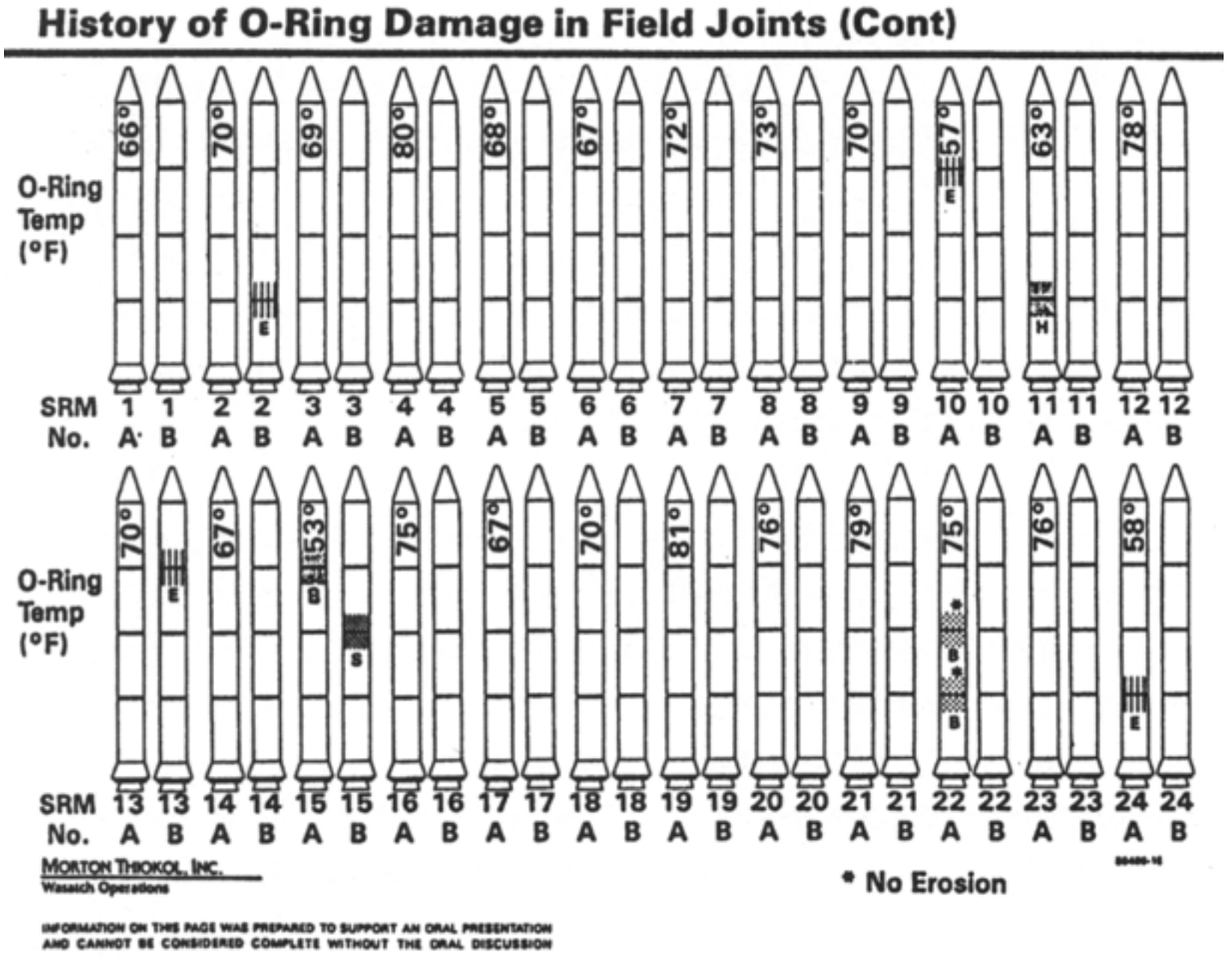
" THIS MAY WELL BE THE WORST GRAPHIC EVER TO FIND ITS WAY TO PRINT. "

- data = five numbers
- area not directly proportional to number
- arbitrary interpolation/smoothing
- a lot of ink!
- colors, 3D & mirroring completely redundant



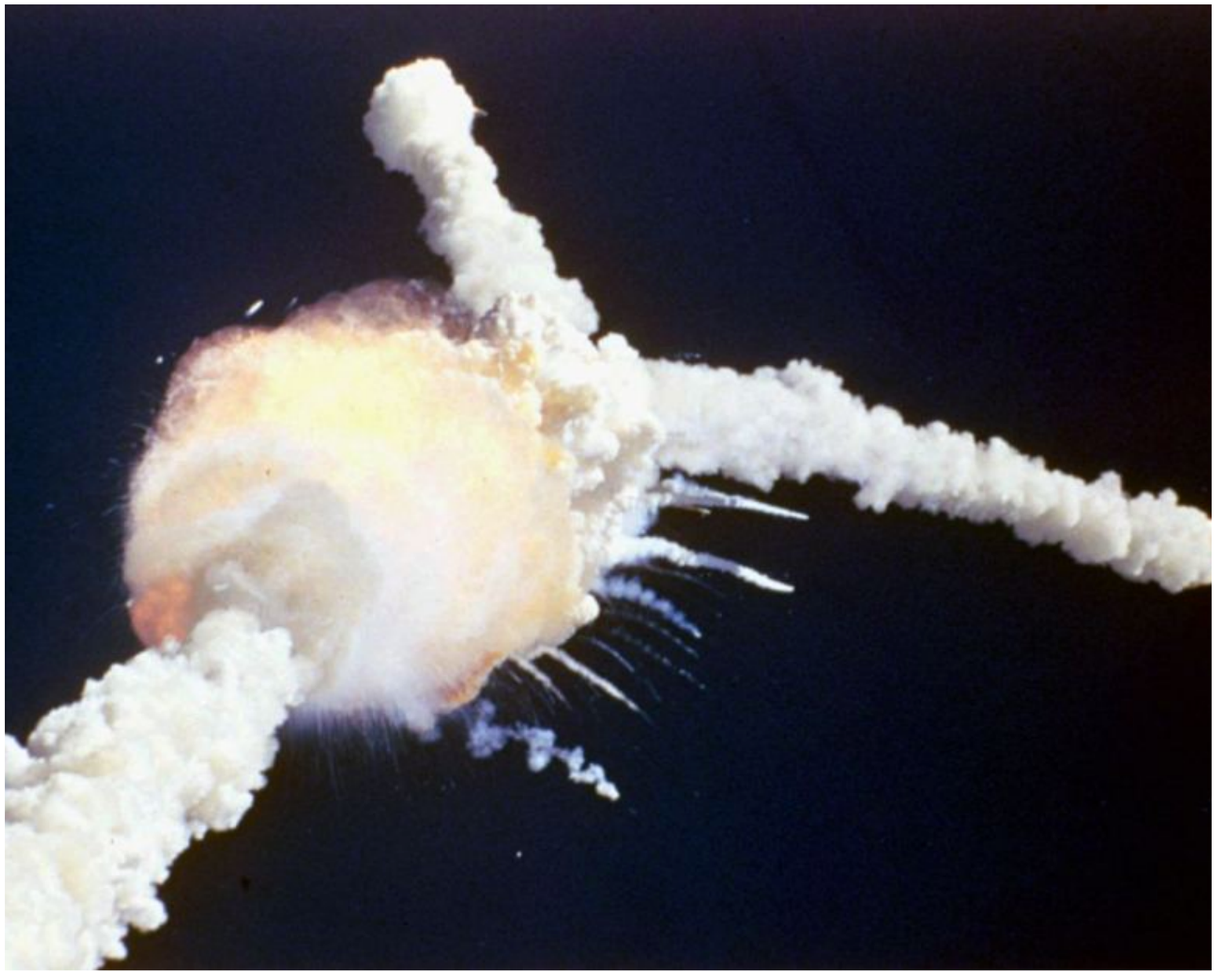
THE GOOD, THE BAD, AND THE UGLY #8

- It's not just about the pretty pictures



THE GOOD, THE BAD, AND THE UGLY #8

- It's not just about the pretty pictures



THE GOOD, THE BAD, AND THE UGLY #8

- It's not just about the pretty pictures

Los Angeles Times

Circulation: 1,076,466 Daily / 1,346,343 Sunday | Tuesday, January 28, 1986 | I,F/ 82 Pages ©Copyright 1986 The Times Mirror Company | Daily 25

Shuttle Explodes; All 7 Die

Teacher on Board as Challenger Blows Up on Liftoff

Reagan Postpones Future Flights Pending a Probe

By MICHAEL SEILER and PETER H. KING,
Times Staff Writers

KENNEDY SPACE CENTER, Fla.—The space shuttl Challenger exploded in a huge fireball less than two minute after takeoff today, with all seven crew members—includin New Hampshire teacher Sharon Christa McAuliffe—feare dead.

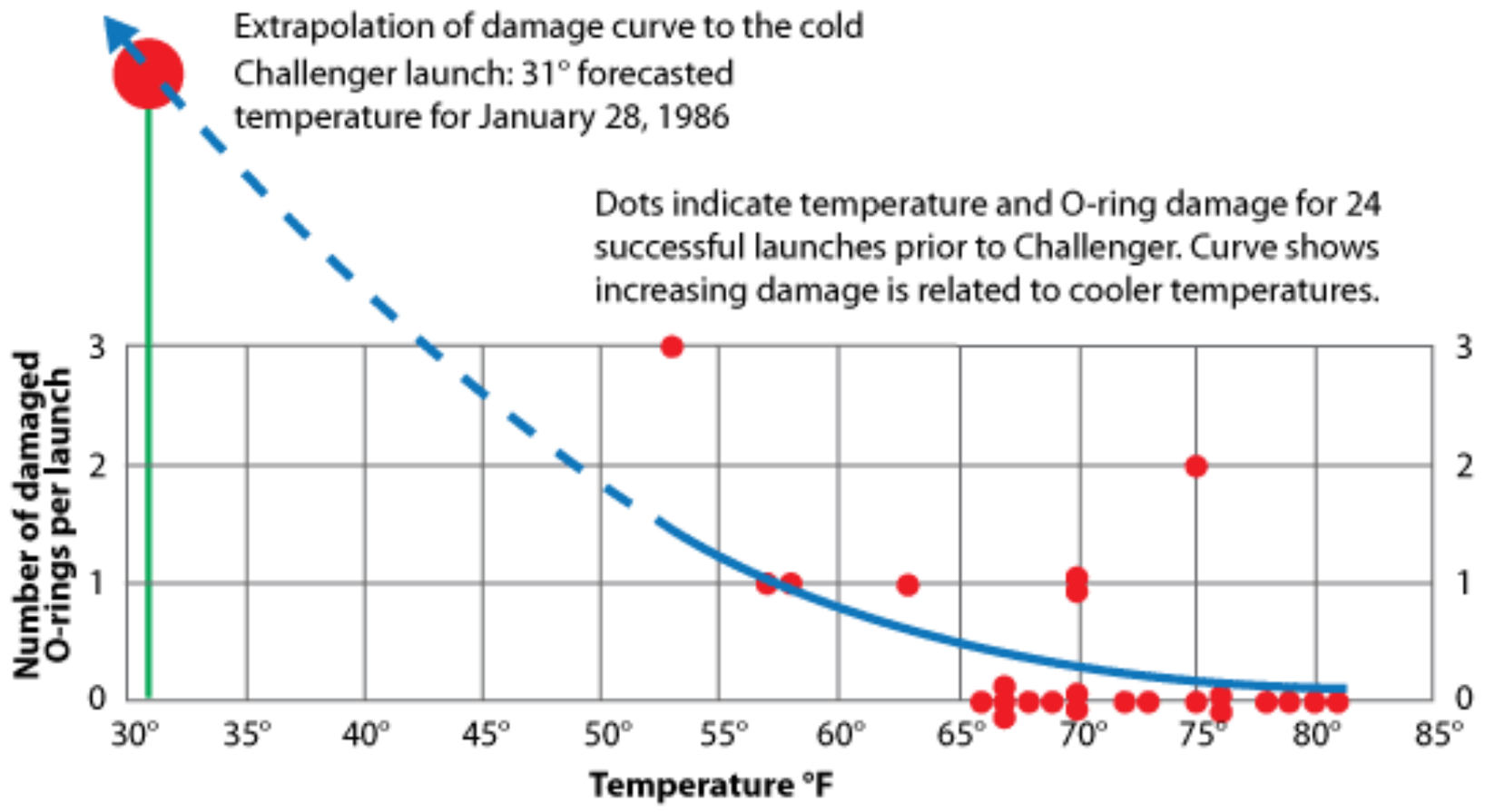
Airborne paramedics parachuted quickly into the cal waters off Cape Canaveral in a vain search for survivor Though there was no immediate announcement on the fate o the crew, all were believed dead.

The disaster—the worst in the history of America's manned space program—came shortly after the Challenger blasted off on a cold Florida morning on the 25th shuttle

As the contrails formed to th east of Cape Canaveral, a parachut appeared in the clear blue sky giving spectators a small glimme

THE GOOD, THE BAD, AND THE UGLY #8

- It's not just about the pretty pictures



ONE MORE TIP

GGPLOT2 PACKAGE IN R

- Transparency or jitter helps with overlapping points

```
> D <- read.csv("player_stats.csv")
> library(ggplot2)
> p = ggplot(D, aes(height, weight))
> p = p + geom_point(color='red',alpha=.03)
> p
```