Principles of Data Visualization

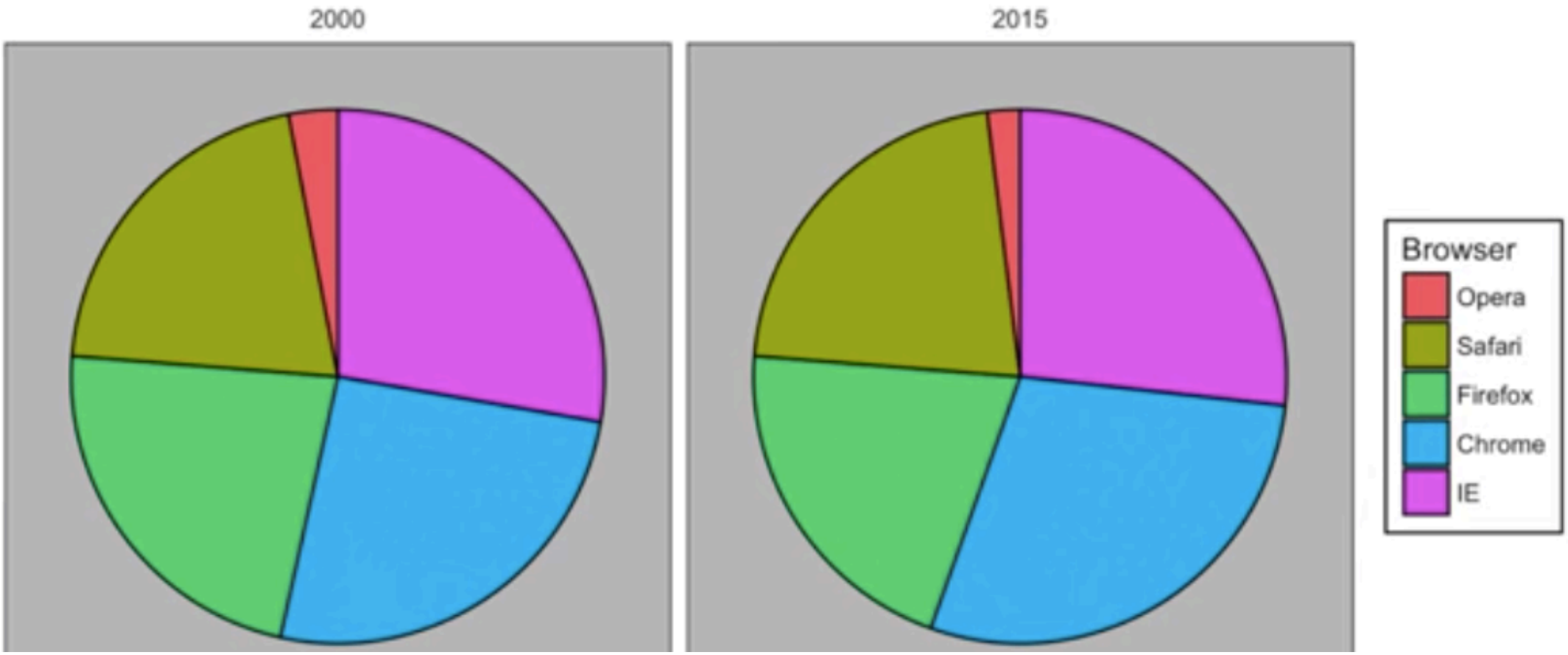
Sumeet Gupta

Encoding Data using Visual Cues

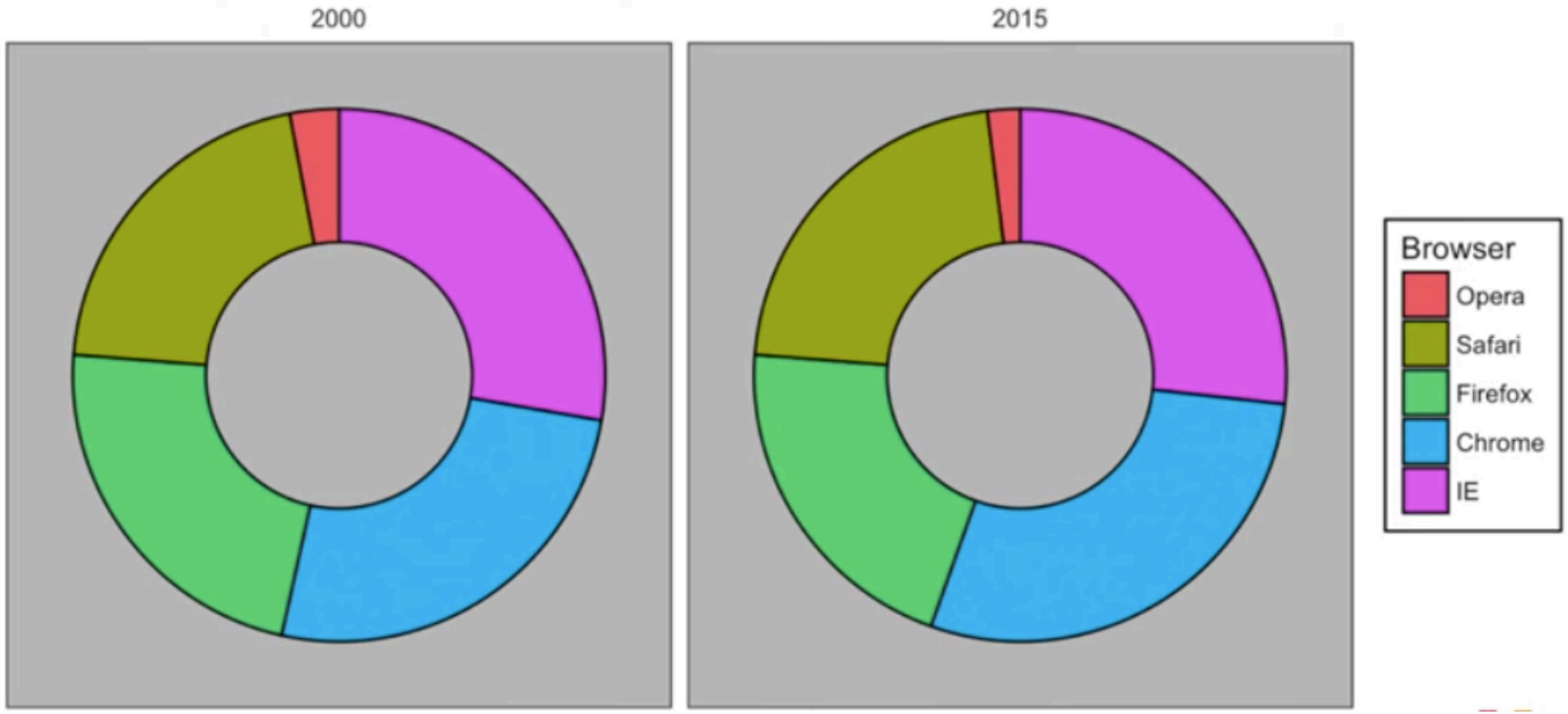
Principle 1

- Position and length are the preferred way to display quantities over angles, which are preferred over area. Brightness and color are even harder to quantify than angles and area. But they are sometimes useful when more than two dimensions are being displayed.

Encoding Data using Visual Cues

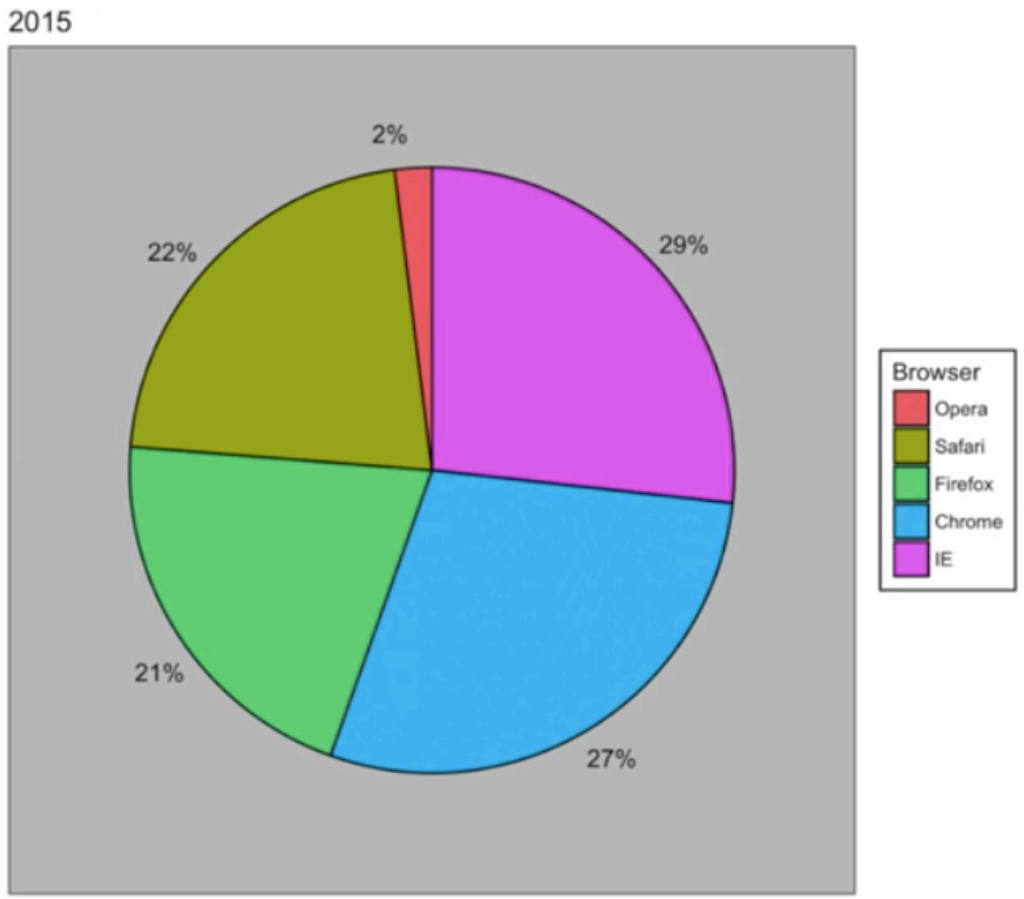


Encoding Data using Visual Cues

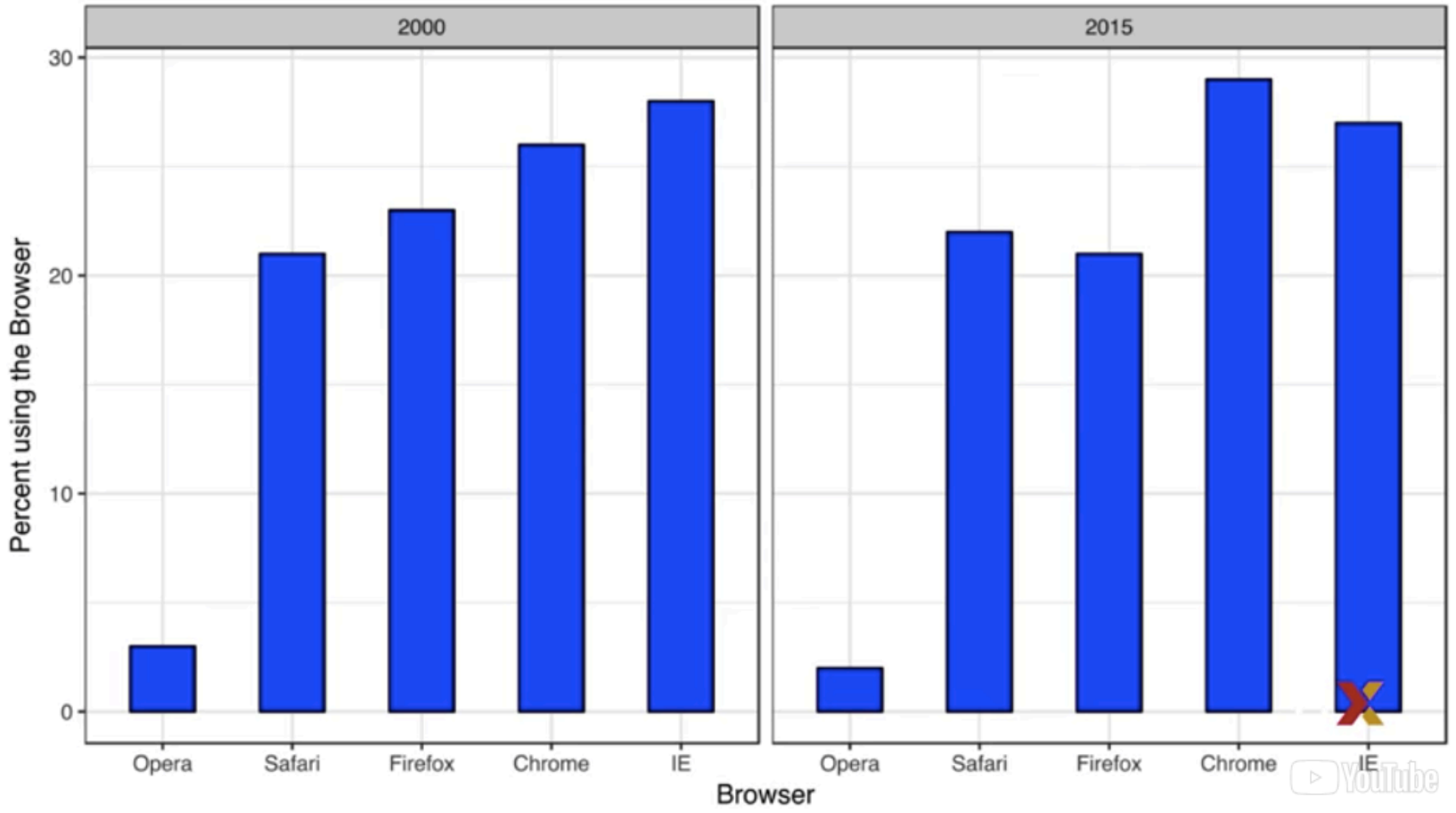


Encoding Data using Visual Cues

Browser	2000	2015
Opera	3	2
Safari	21	22
Firefox	23	21
Chrome	26	29
IE	28	27



Encoding Data using Visual Cues

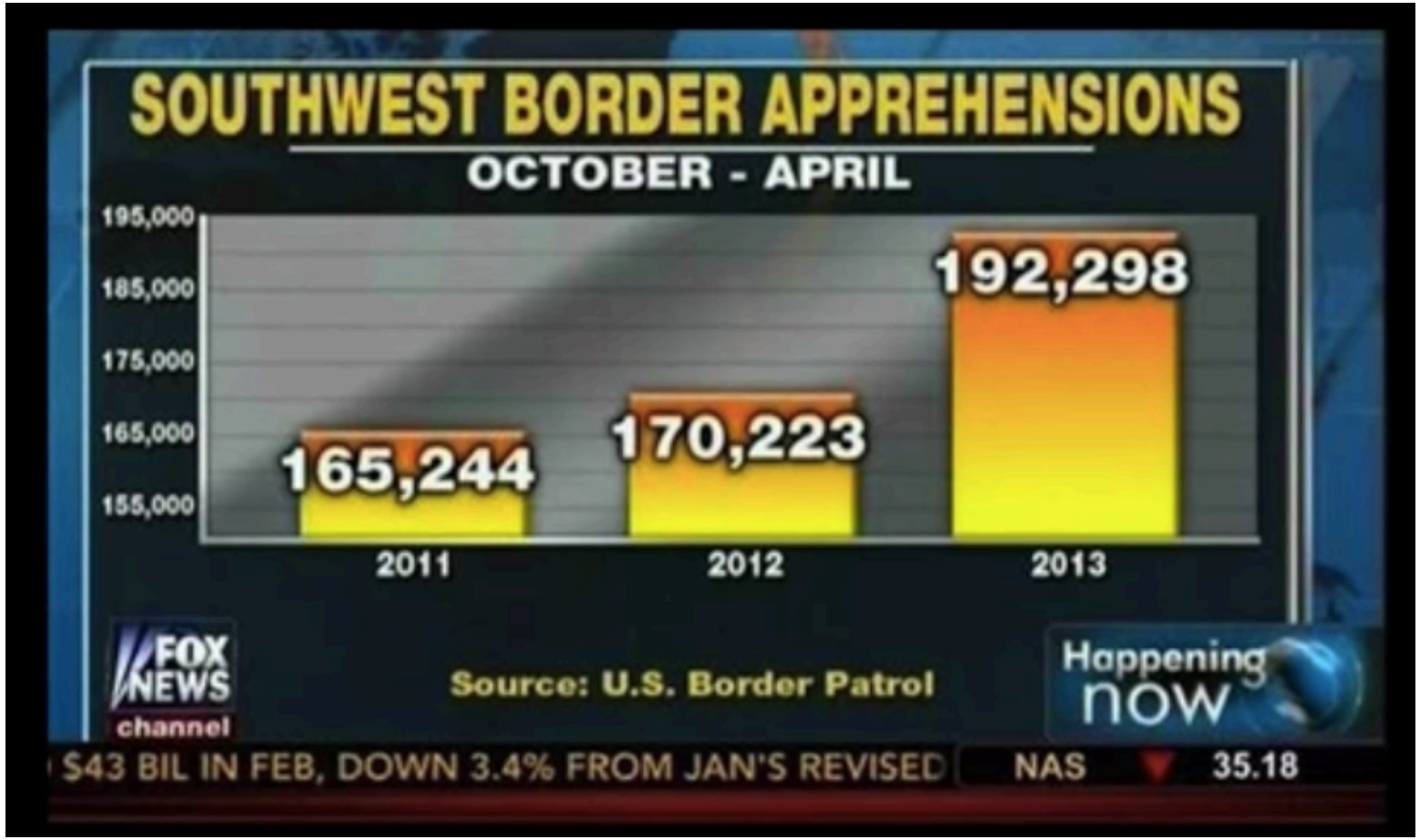


Know when to include zero

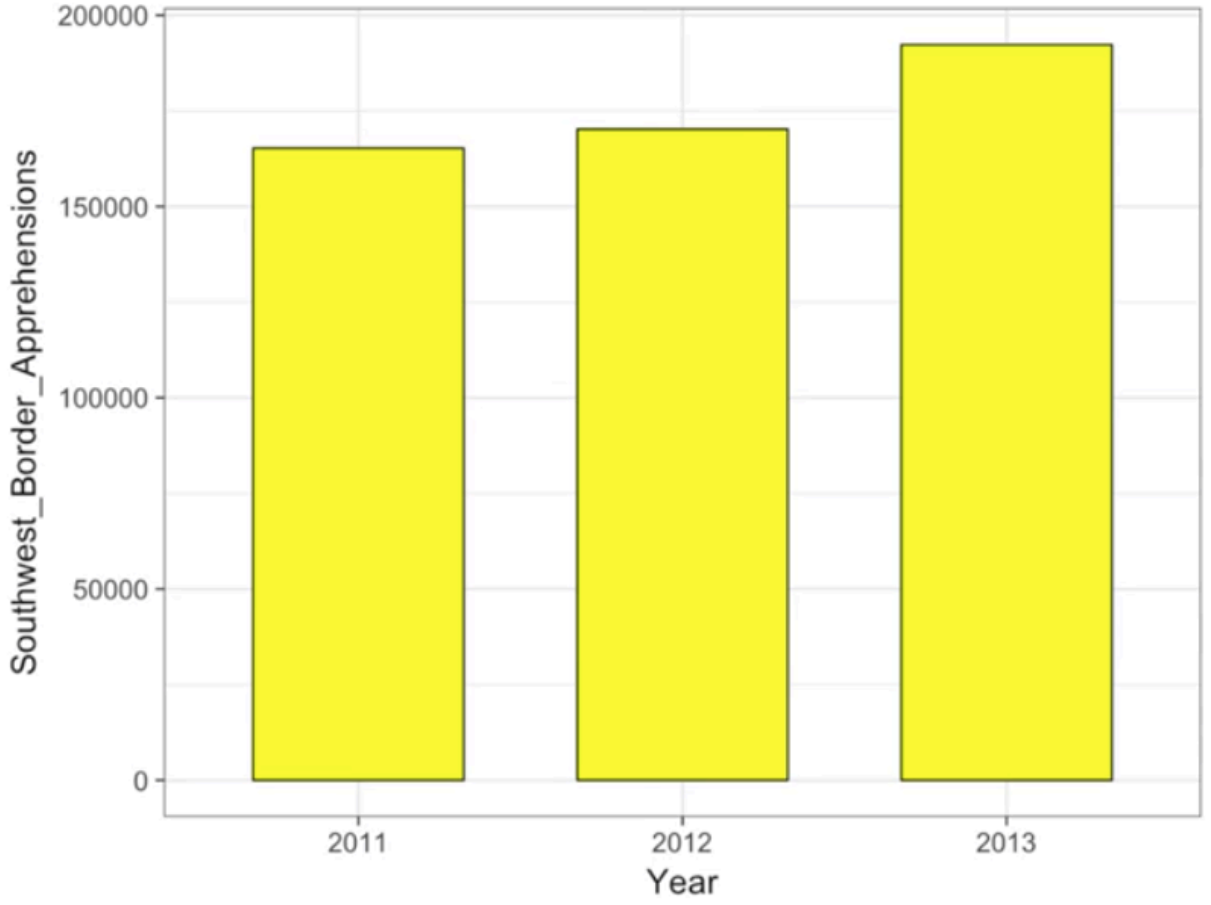
Principle 2

- When using bar plots, it is dishonest not to start the bars at 0. This is because by using a bar plot, we are implying the length is proportional to the quantities being displayed. By avoiding 0, relatively small differences can be made to look much bigger than they actually are.

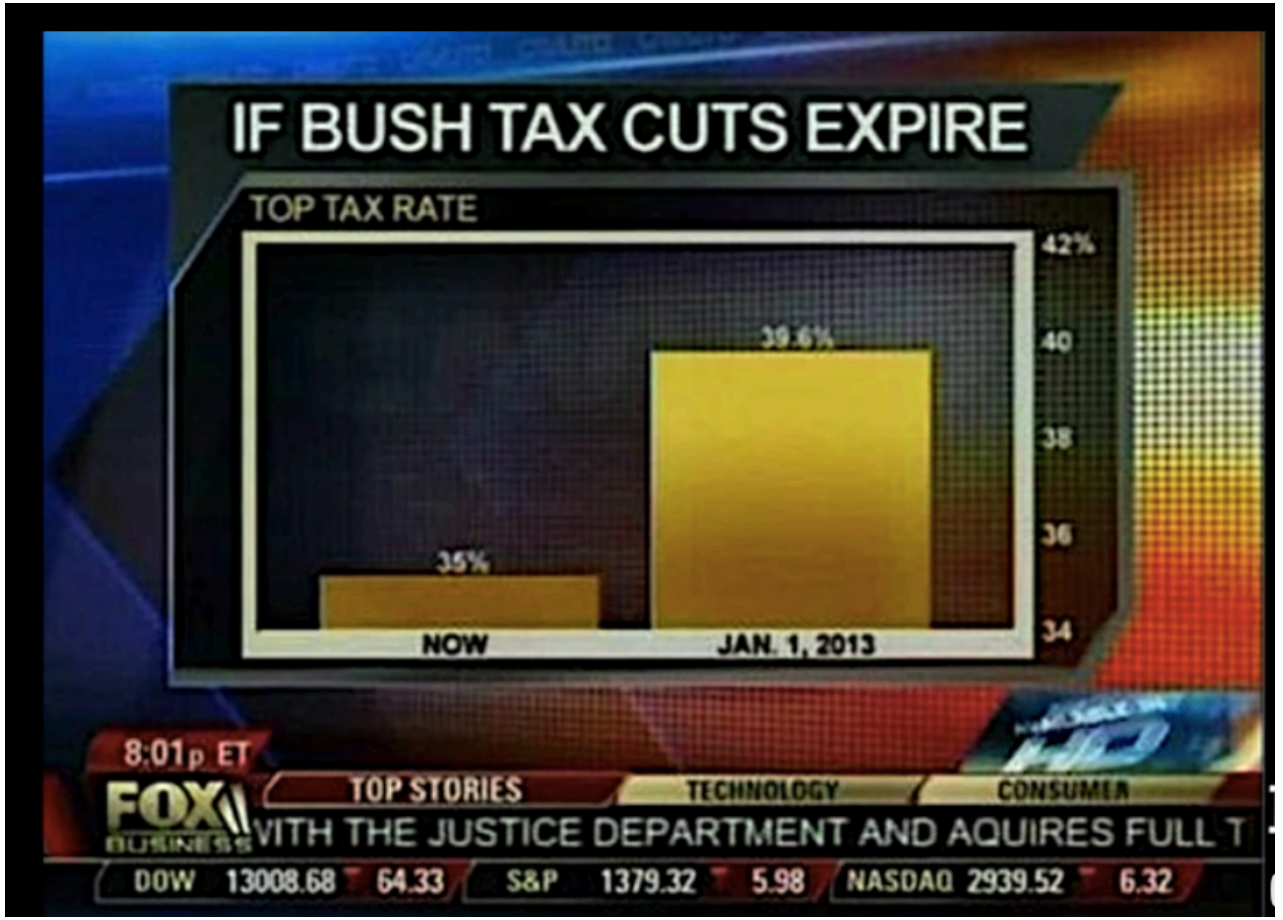
Know when to include zero



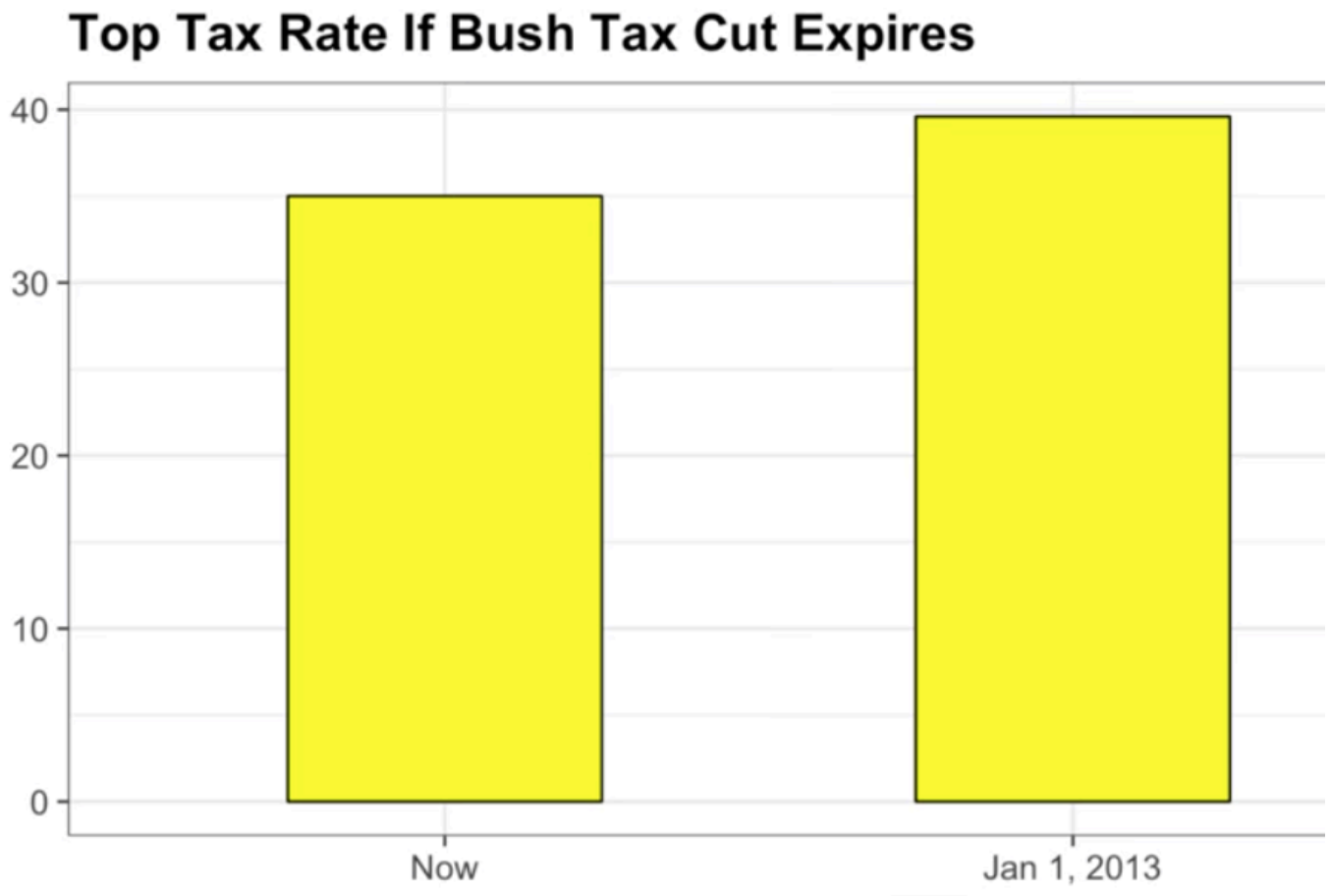
Know when to include zero



Know when to include zero



Know when to include zero

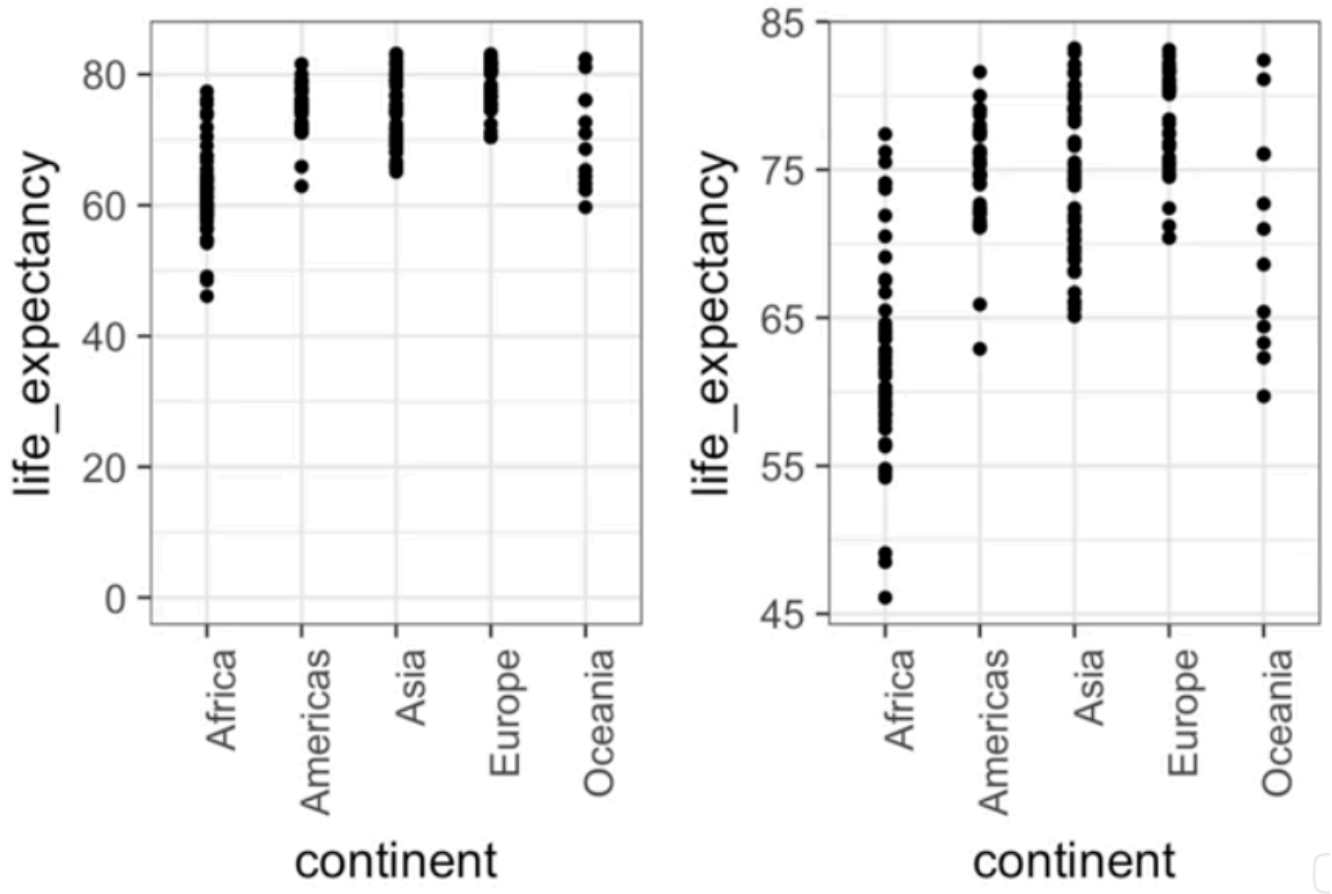


Know when to include zero

Principle 3

- When using position rather than length, it's not necessary to include 0. This is particularly the case when we want to compare differences between groups relative to the variability seen within the groups.

Know when to include zero

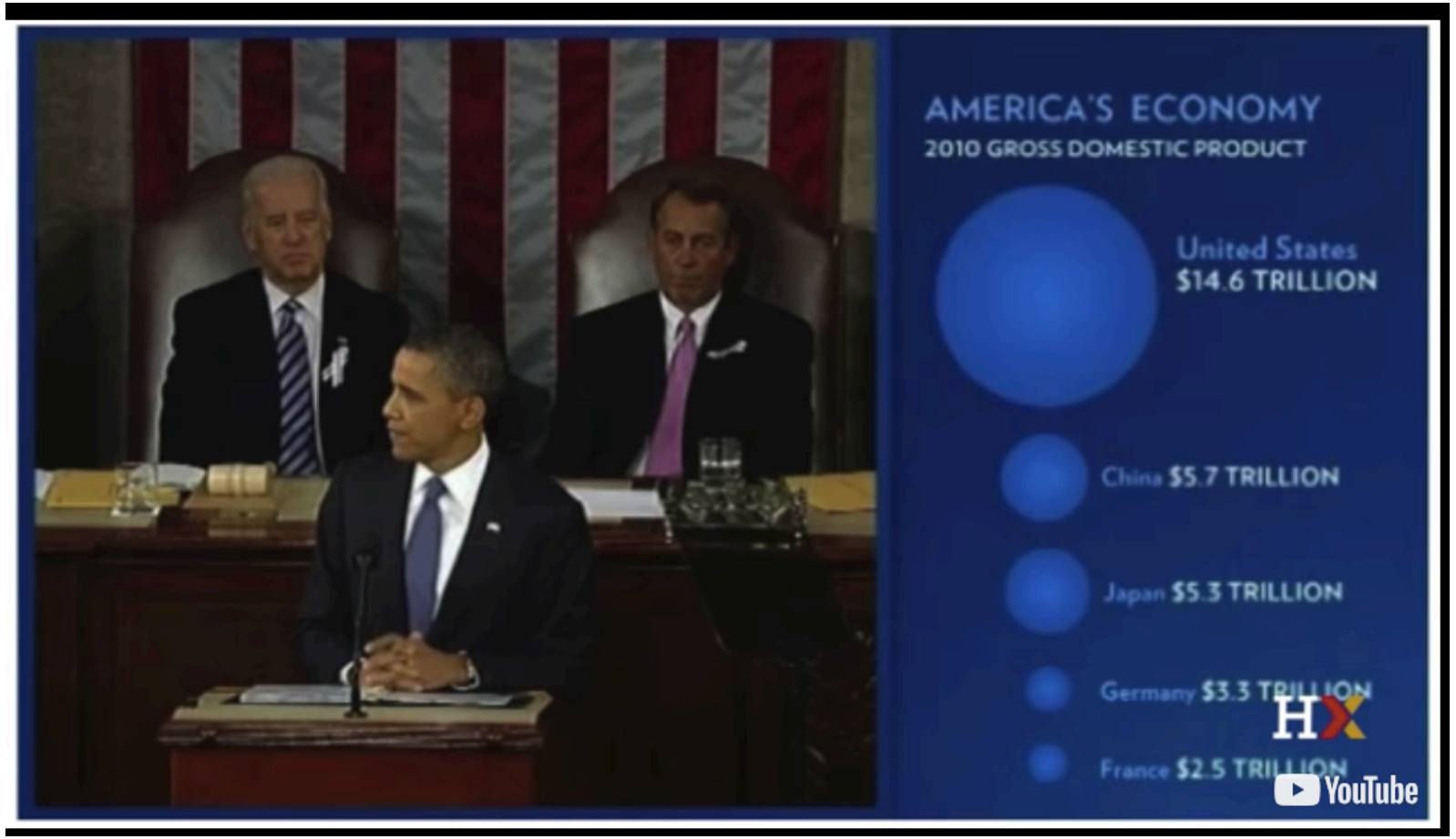


Do not distort quantities

Principle 4

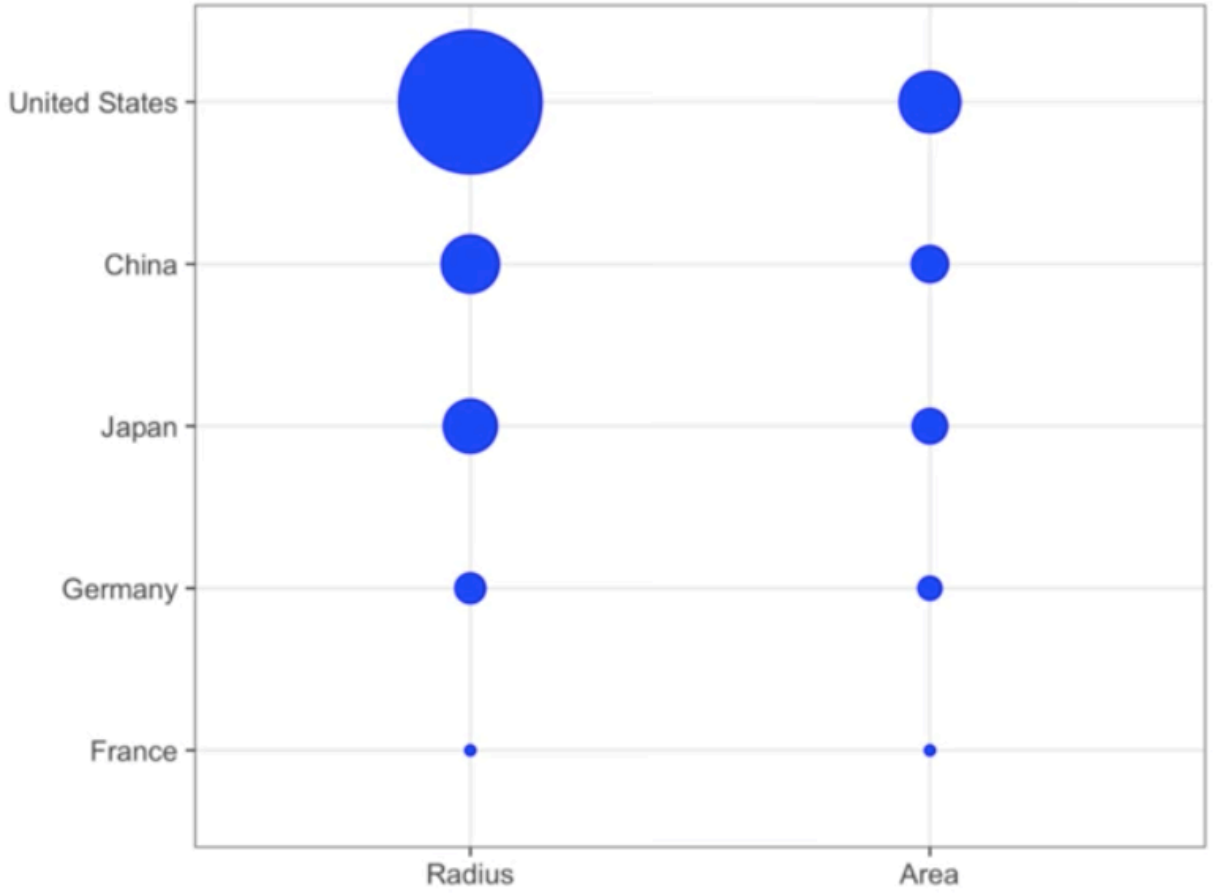
- Do not distort quantities
  - For eg., using radius of a circle for comparing various quanitities than area of the circle

Do not distort quantities

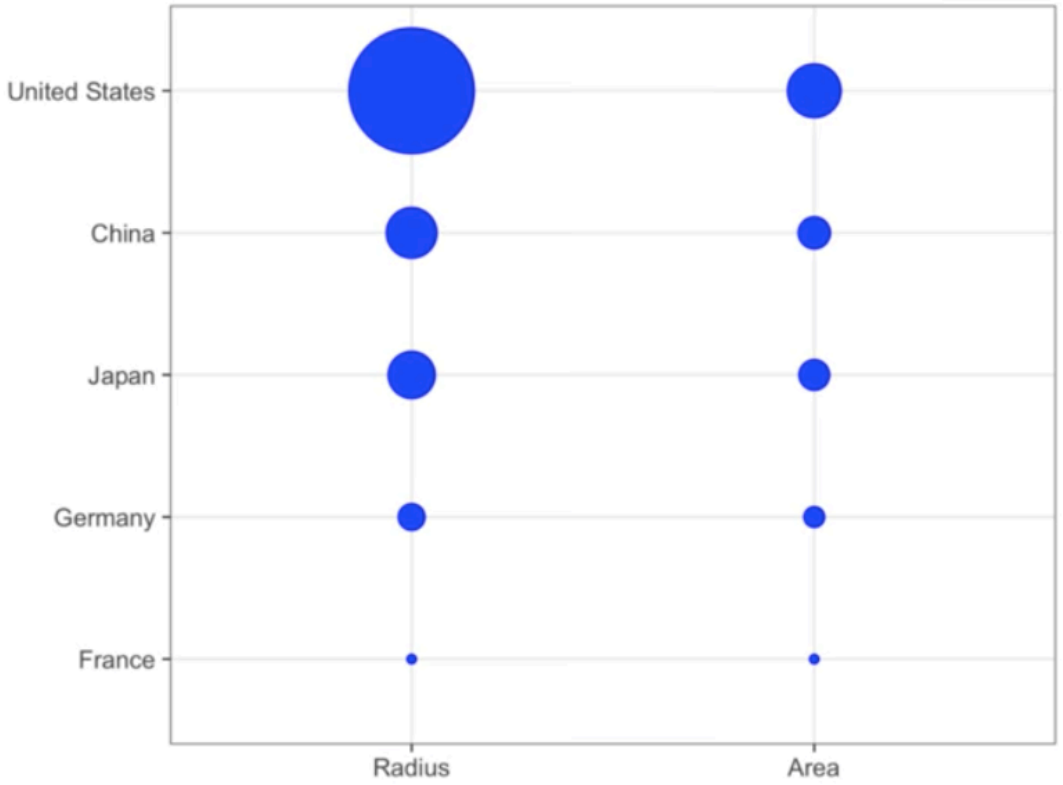


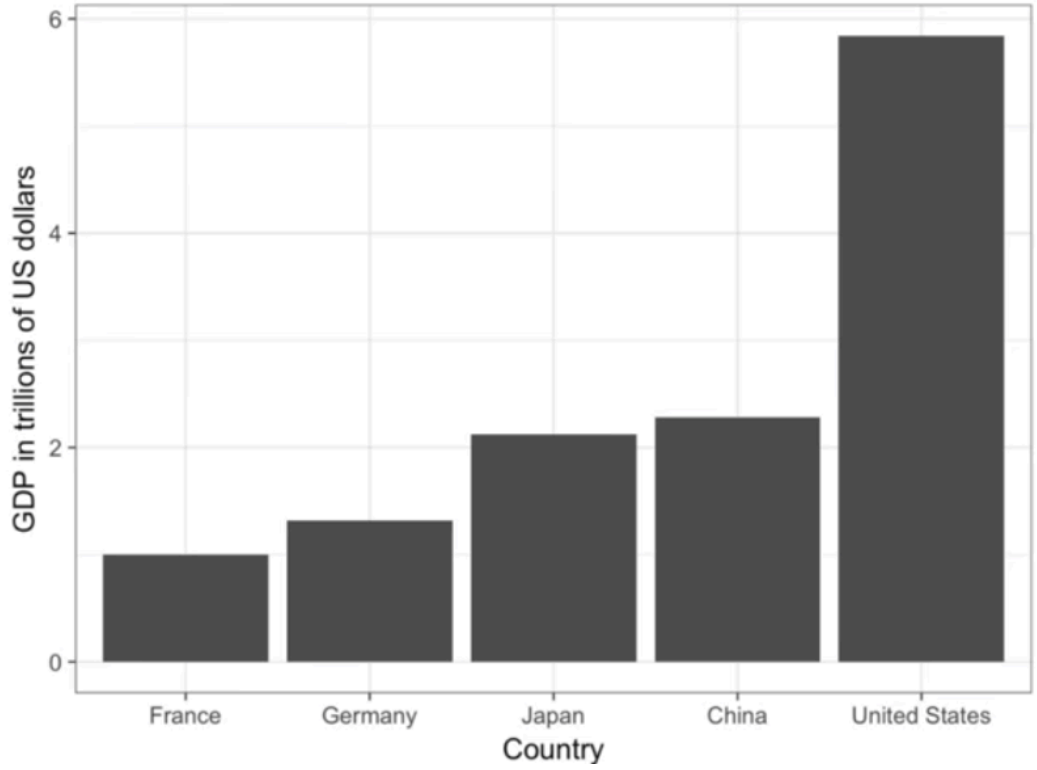
The reason for this distortion is that the radius, rather than the area, was made to be proportional to the quantity, which implies that the proportions between the areas is squared.

Do not distort quantities



Do not distort quantities



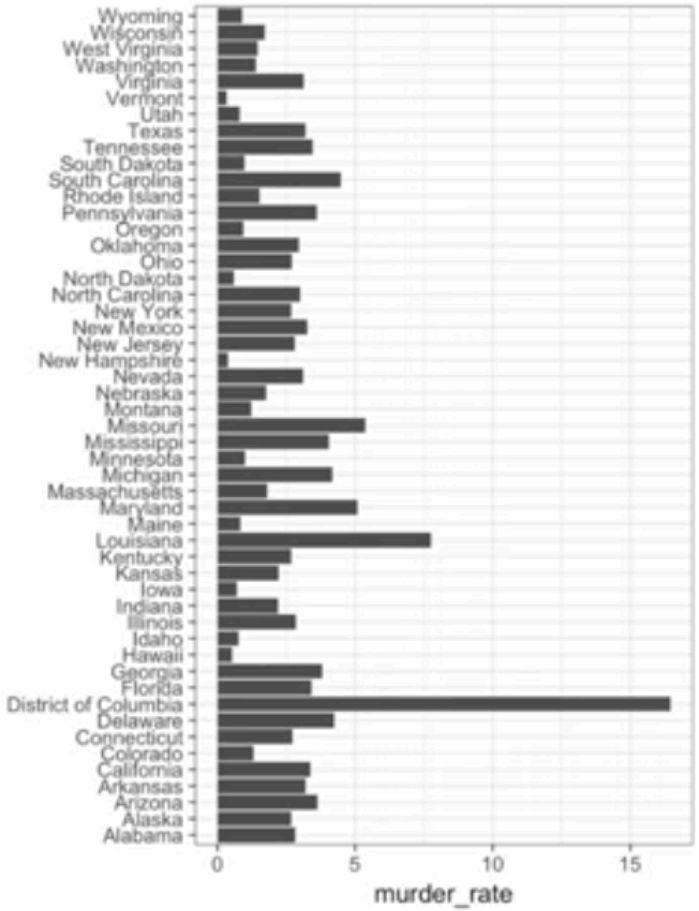


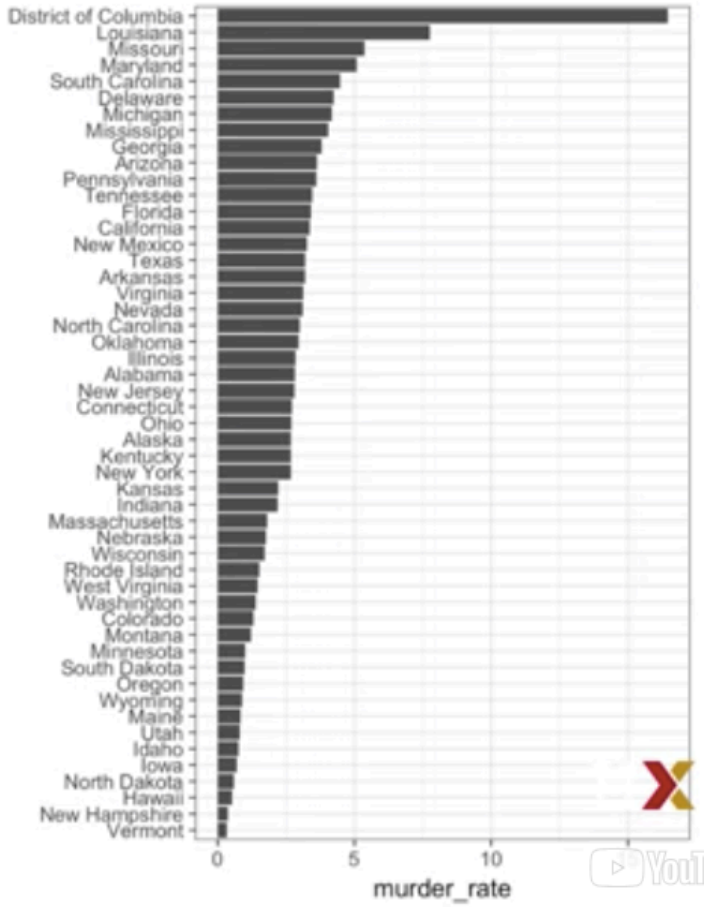
Order by a meaningful value

Principle 5

- When one of the axes is used to show categories, as done in bar plots, the default ggplot behavior is to order the categories alphabetically when they are defined by character strings. Instead, we should order by a meaningful quantity – such as the values being displayed.

Order by a meaningful value

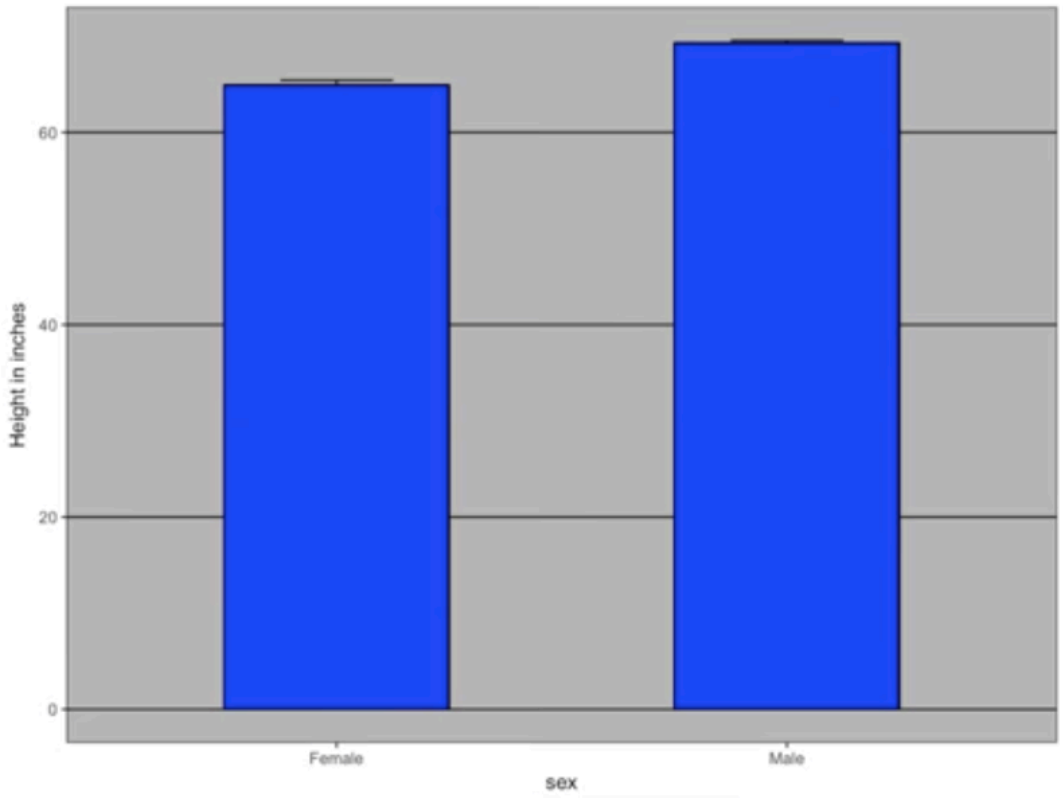




Show the Data

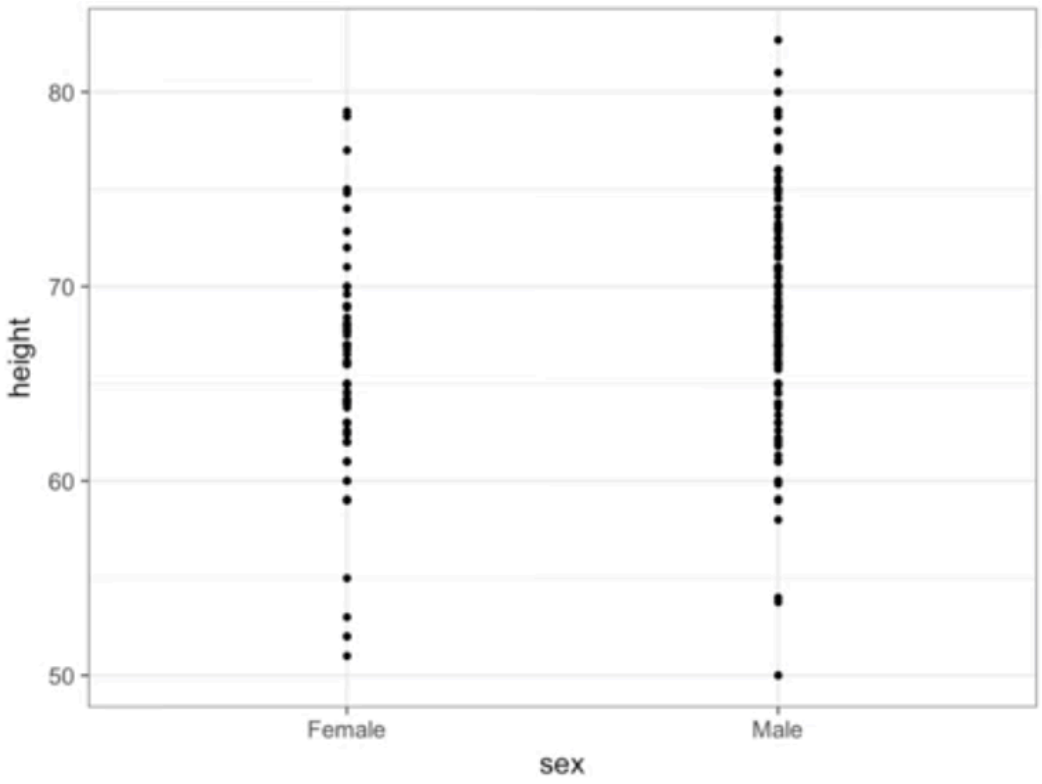
Principle 6

Show the Data

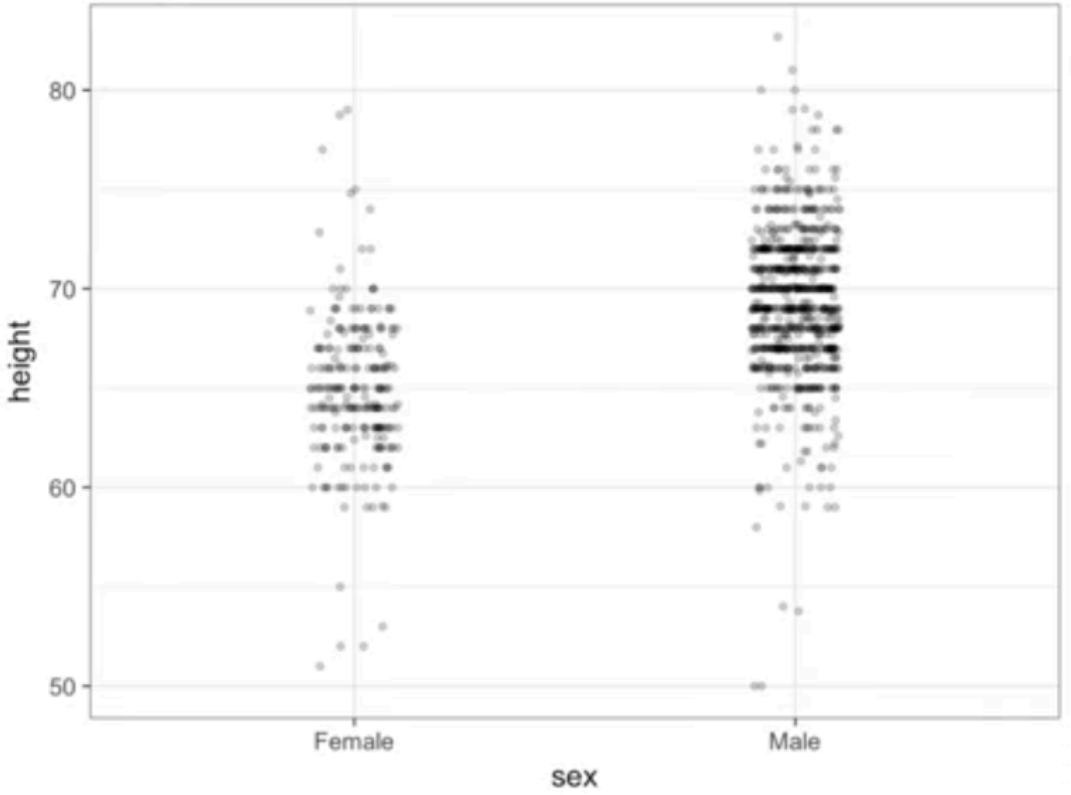


If all ET receives is this plot, he will have little information on what to expect if he meets a group of humans, males and females. Note that the bars go to 0. Does this mean there are tiny humans measuring less than one foot? Are all males taller than the tallest female? Is there a range of heights? One can't answer these questions since we have provided almost no information on the height distribution.

Show the Data



With jitter and alpha blending



Ease Comparisons: Use Common Axis

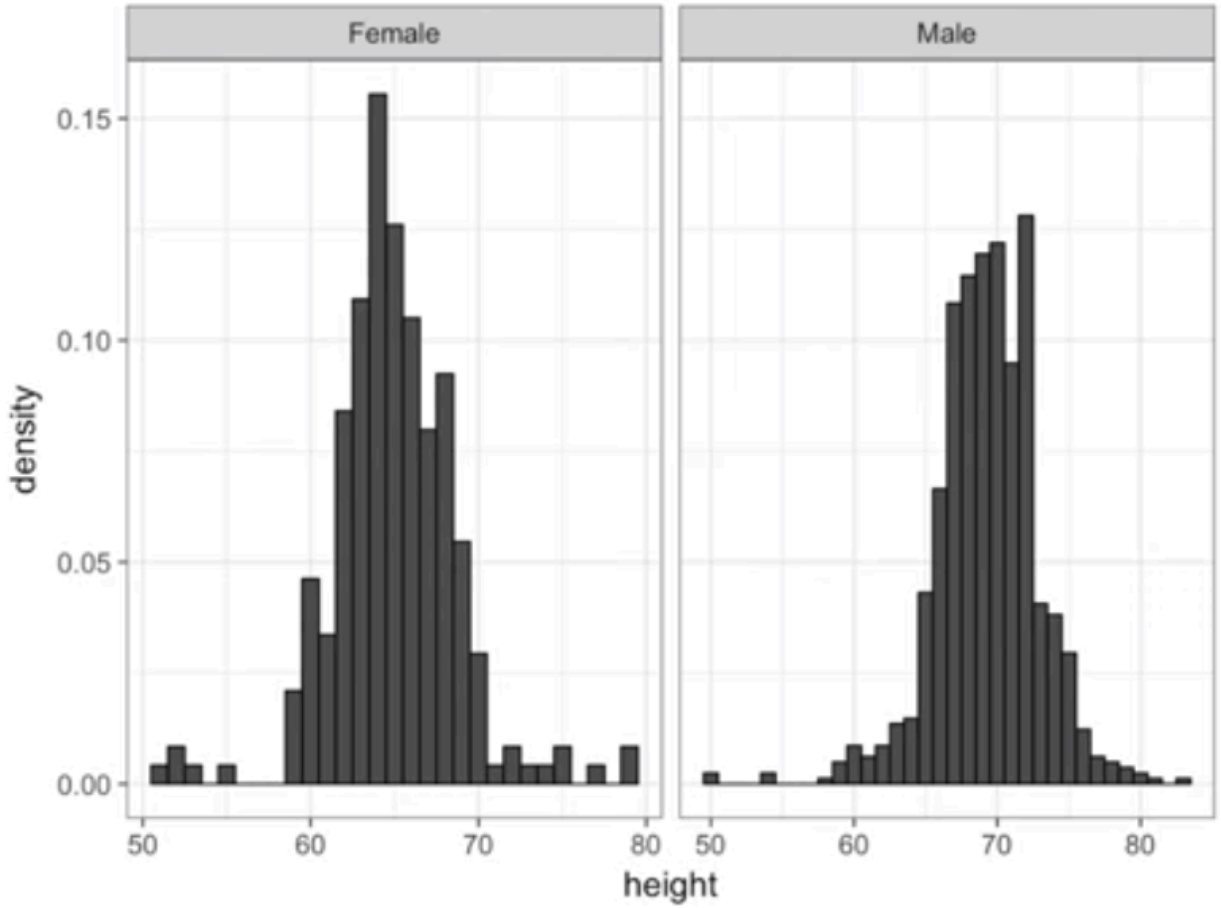
Principle 7

Keep the axes the same when comparing data across plots.

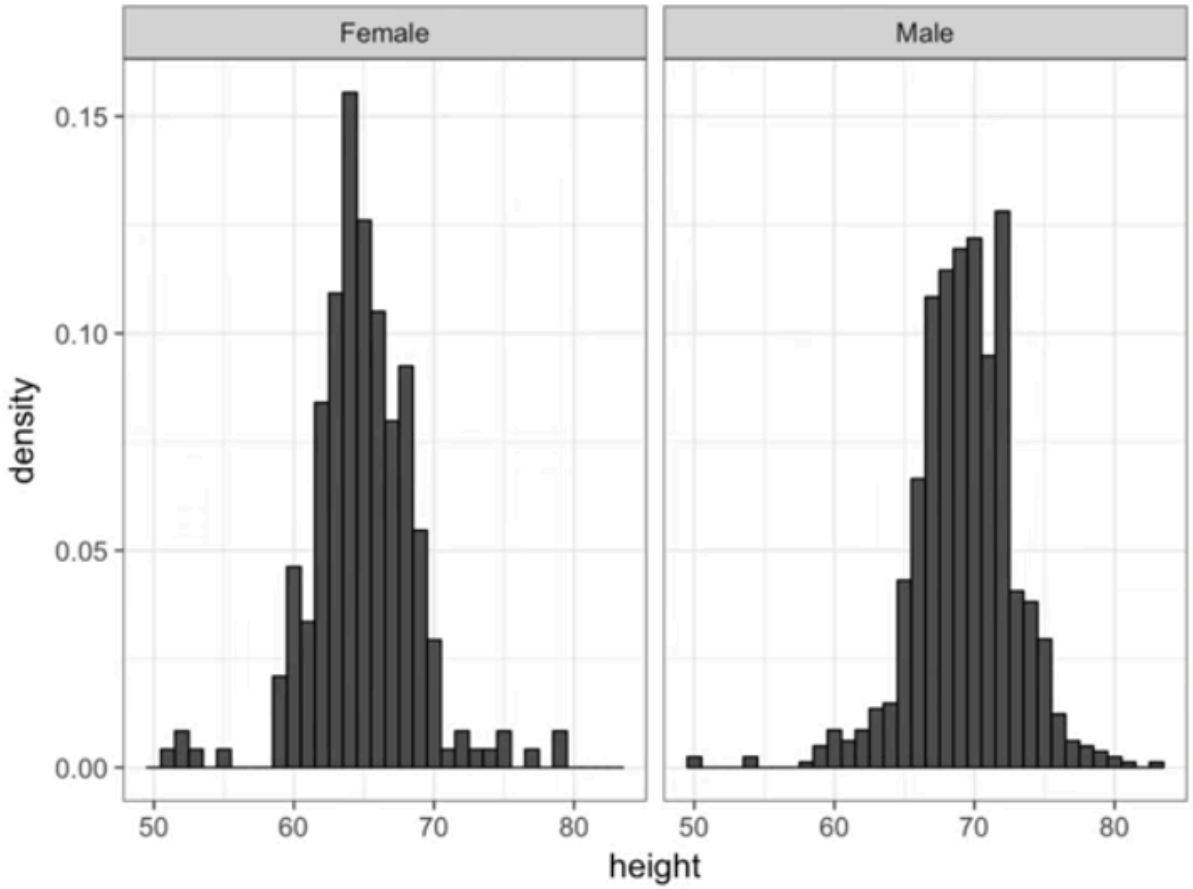
Align plots vertical to see horizontal changes, and horizontally to see vertical changes.

Bar plots are useful for showing one number, but not very useful when wanting to describe distributions.

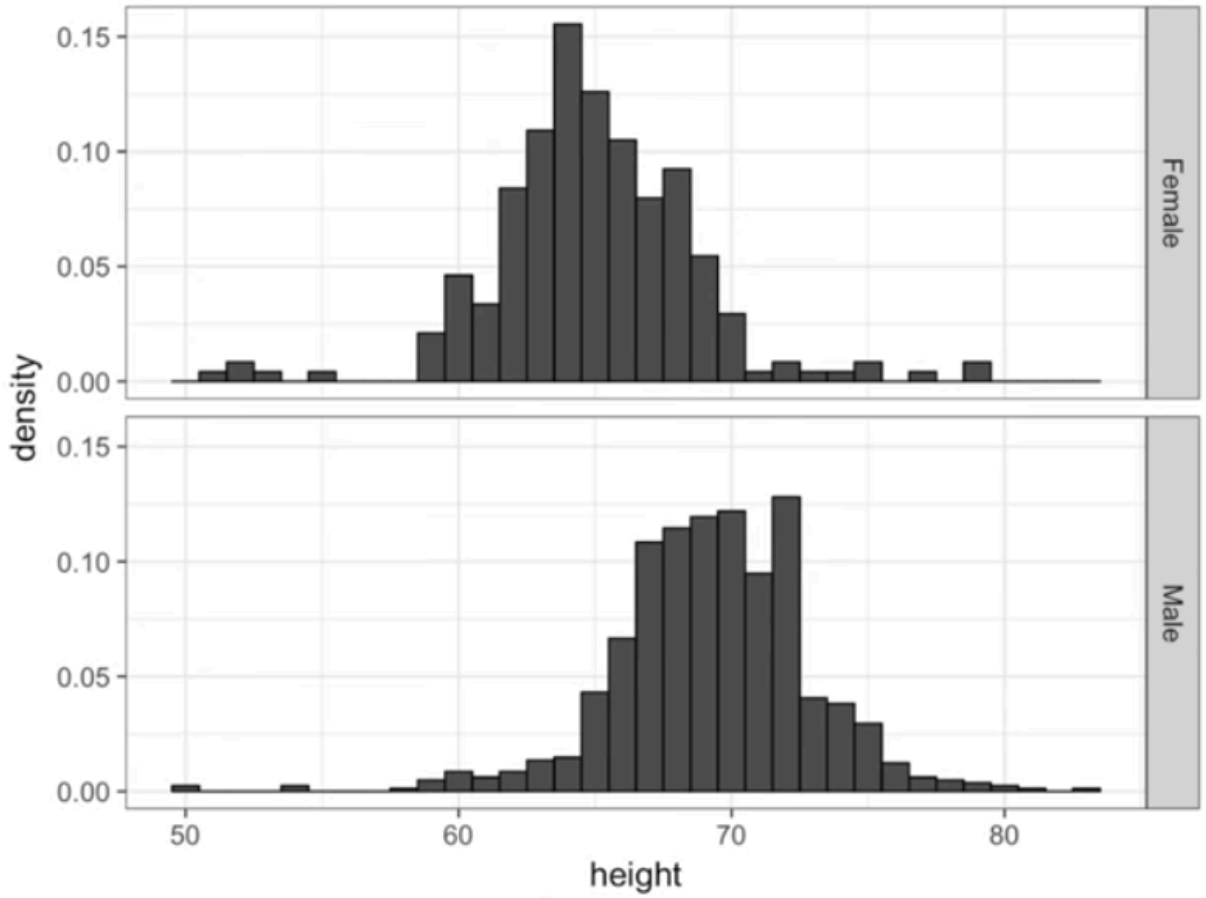
Ease Comparisons: Use Common Axis



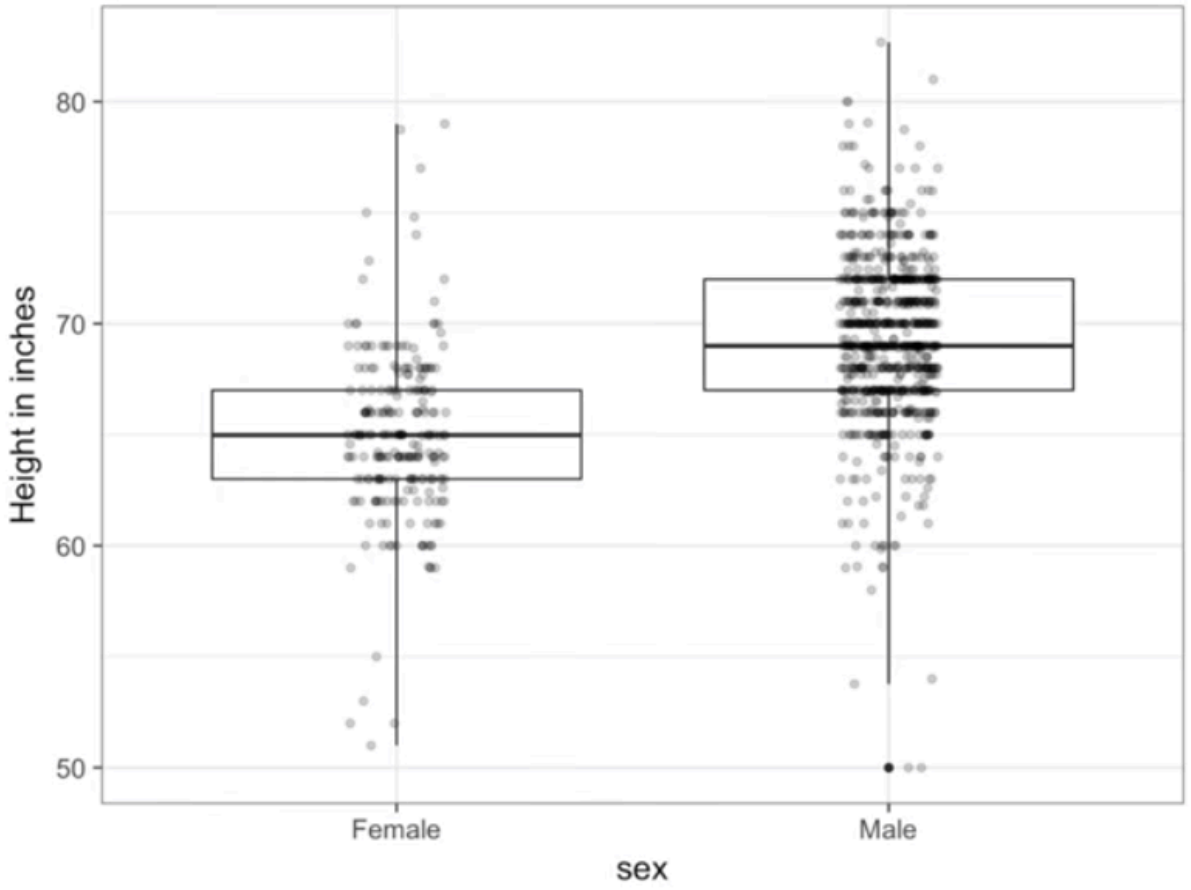
Ease Comparisons: Use Common Axis



Ease Comparisons: Use Common Axis



Ease Comparisons: Use Common Axis

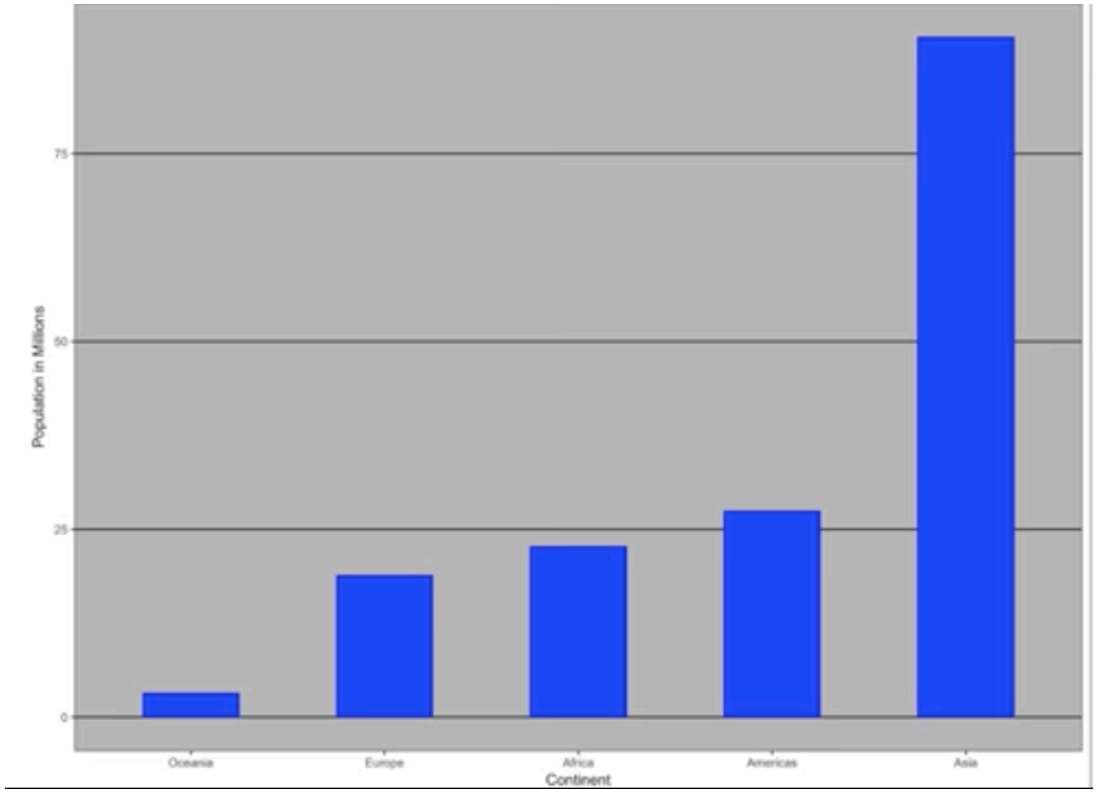


Consider Transformations

Principle 9

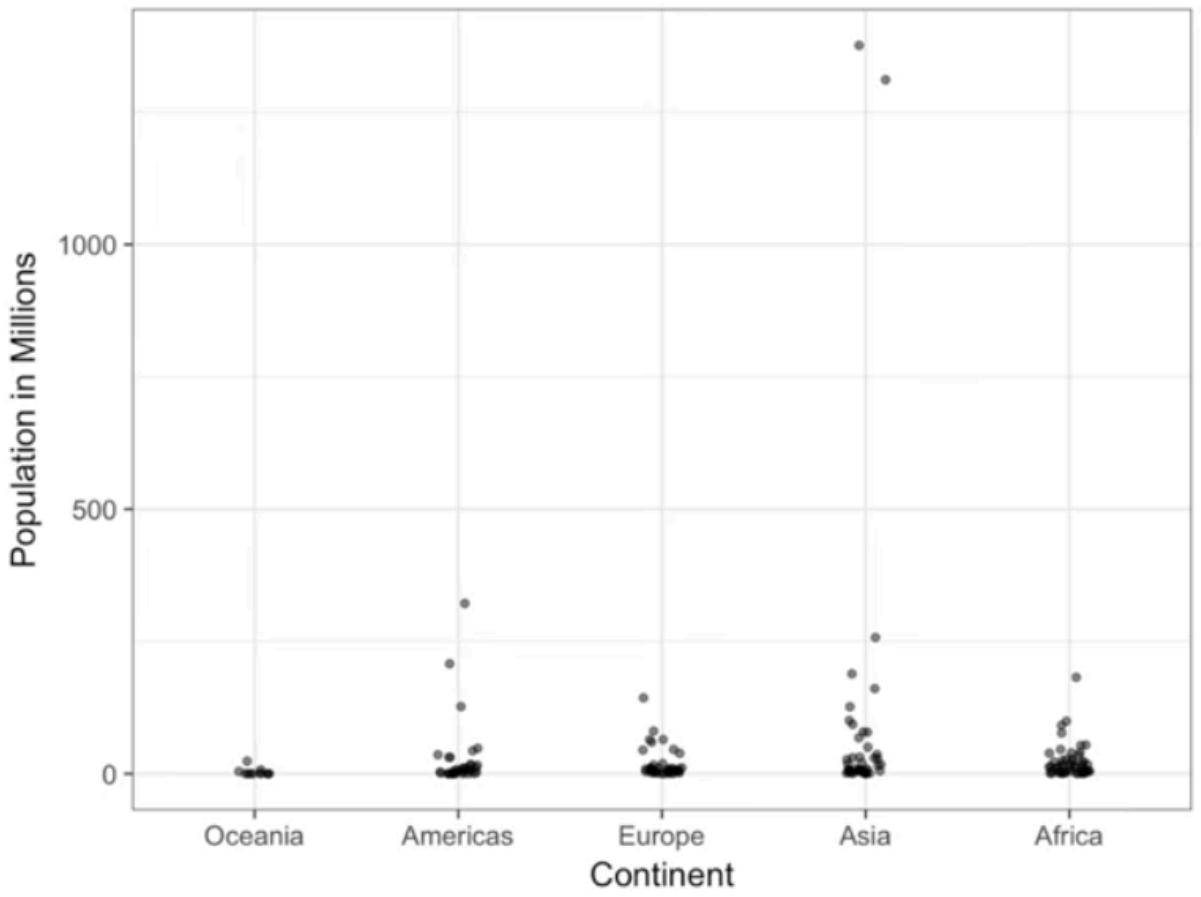
- Use the log transformation in cases where the changes are multiplicative.
- Other useful transformations
  - Logistic transformation useful to better see changes in odds
  - Square Root transformation useful for count data

Consider Transformations

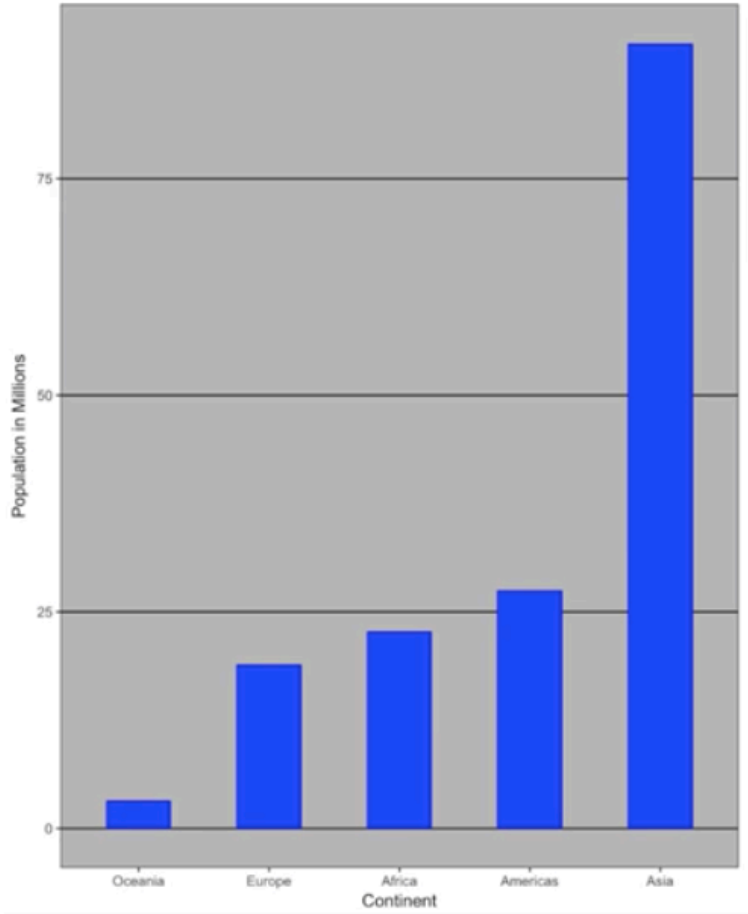


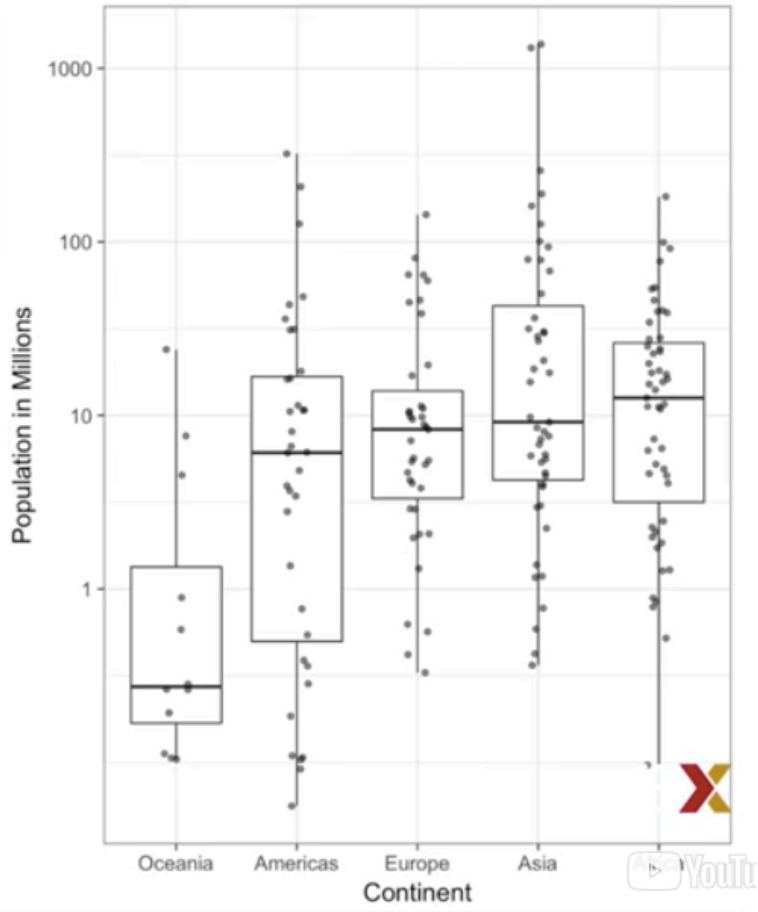
Average population sizes for each continent in 2015.

Consider Transformations



Consider Transformations

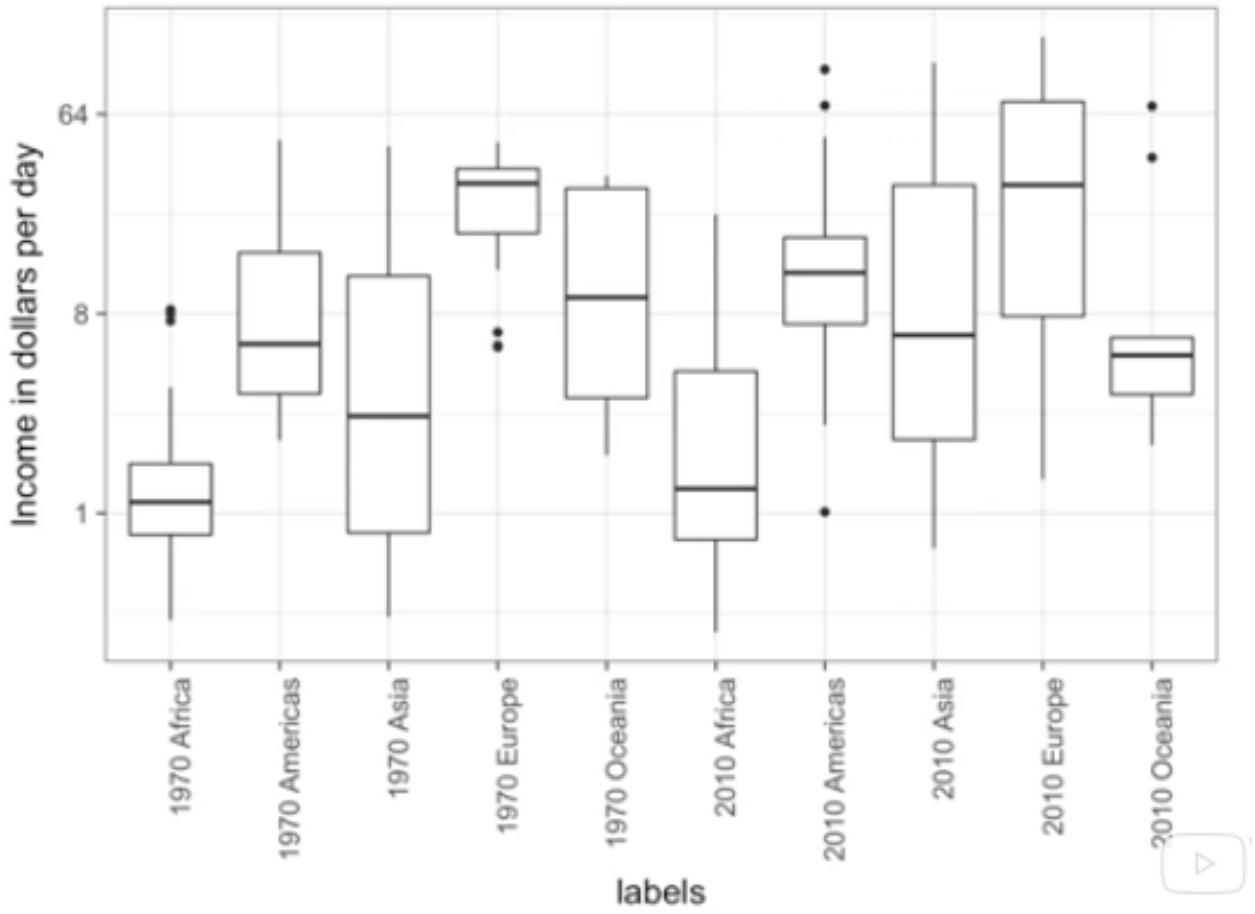




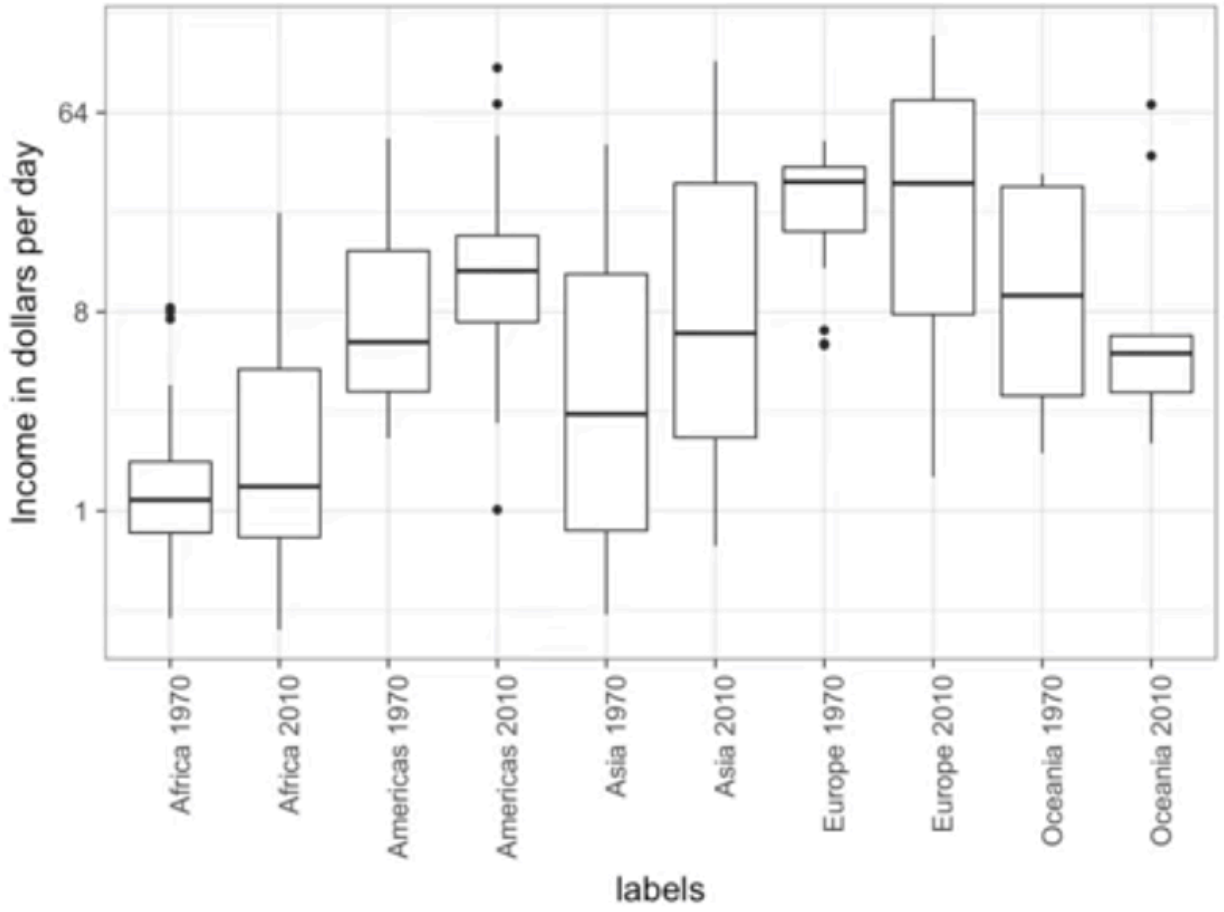
Ease Comparisons-Compared visual cues should be adjacent

Principle 10

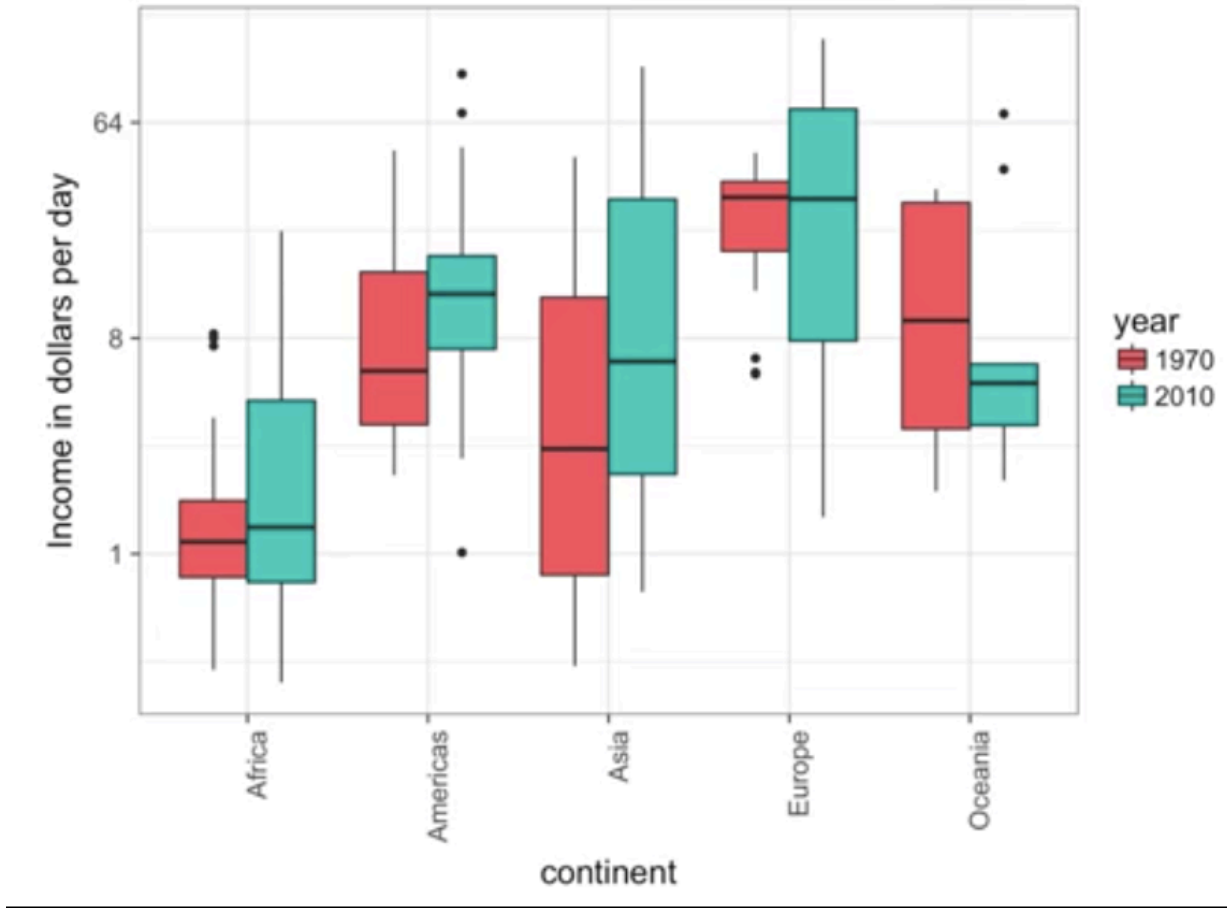
Ease Comparisons-Compared visual cues should be adjacent



Ease Comparisons-Compared visual cues should be adjacent



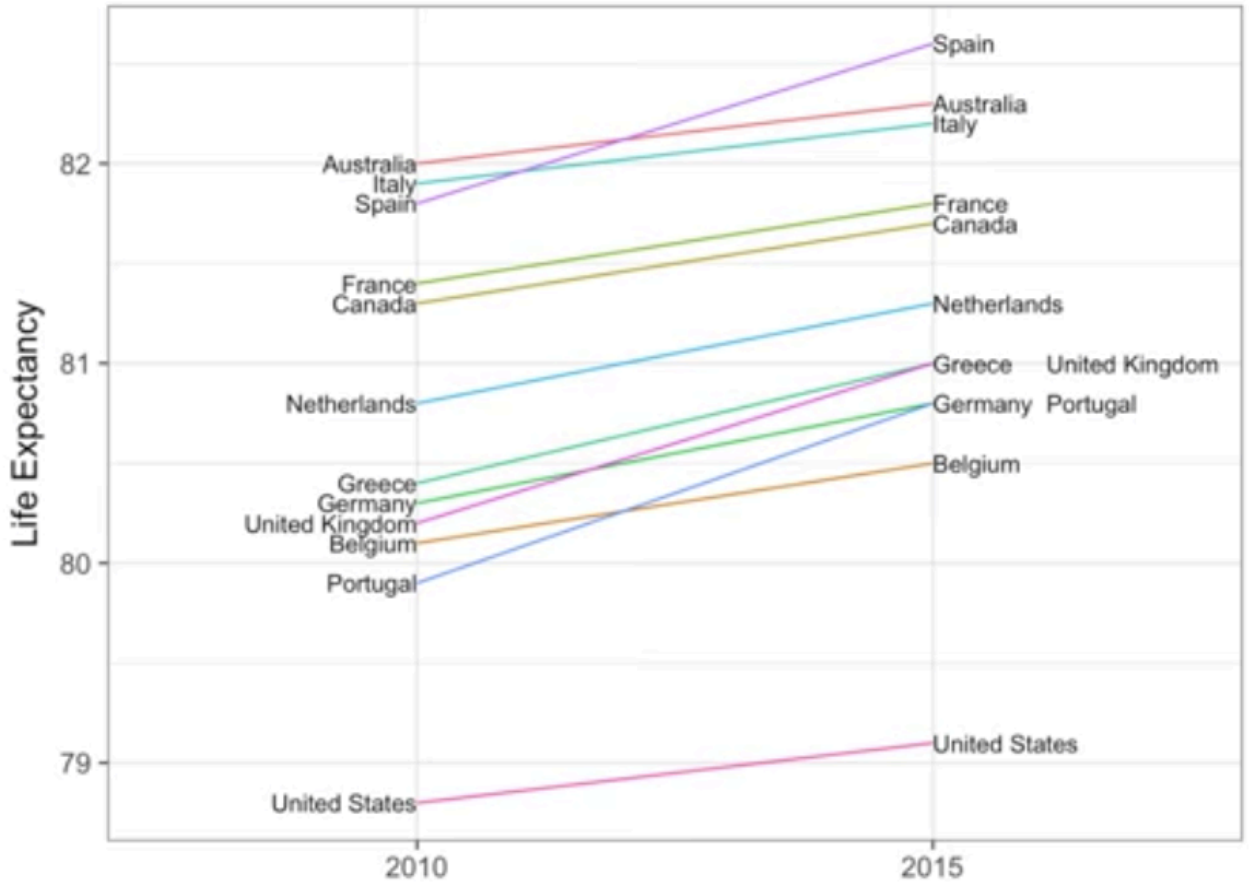
Ease Comparisons-Compared visual cues should be adjacent



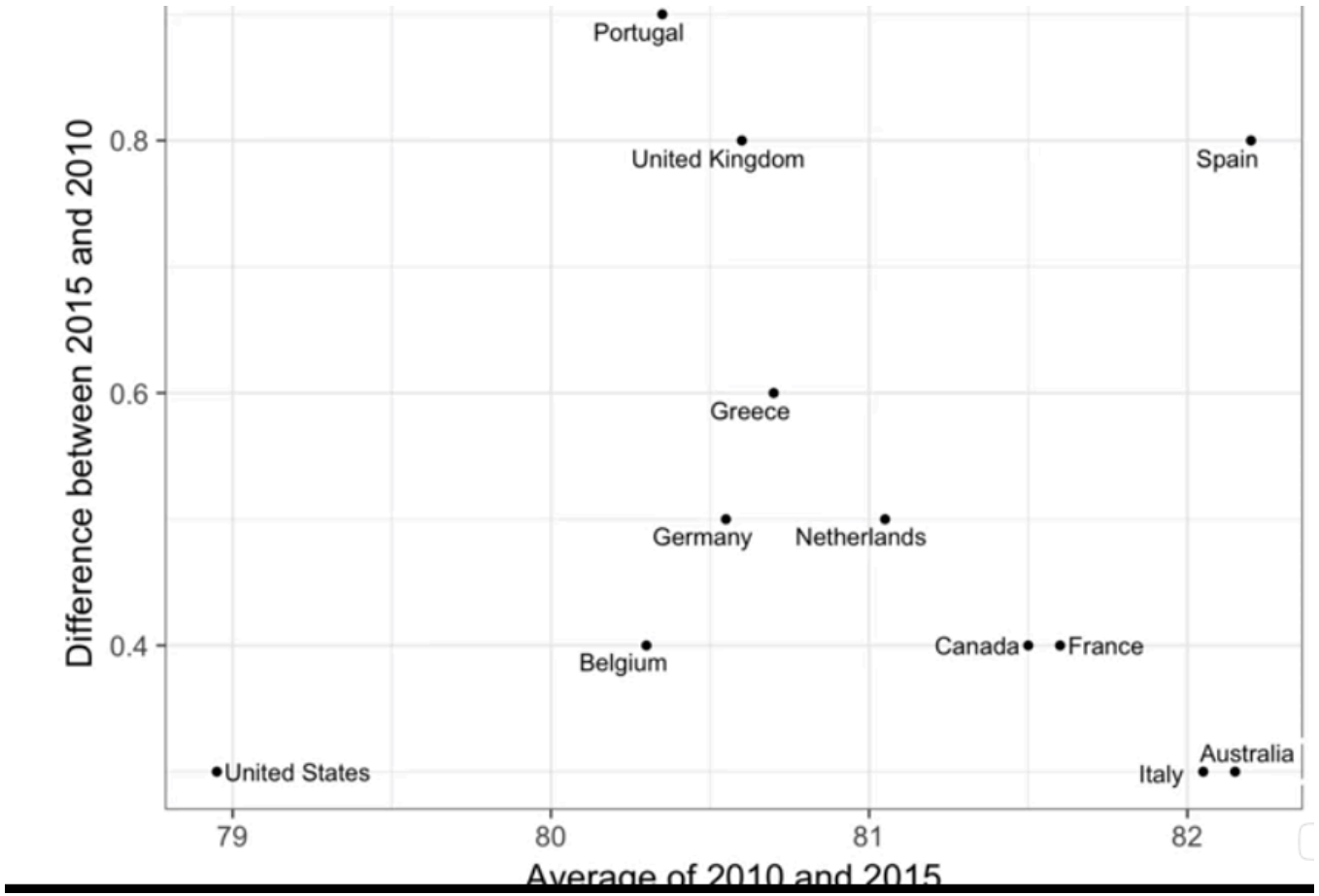
Slope Charts

- **Principle 11**
- Slope charts could be preferred over scatter plots for two variables when comparing variables of the same type but at different time points and for a relatively small number of comparison.

Slope Charts



Slope Charts



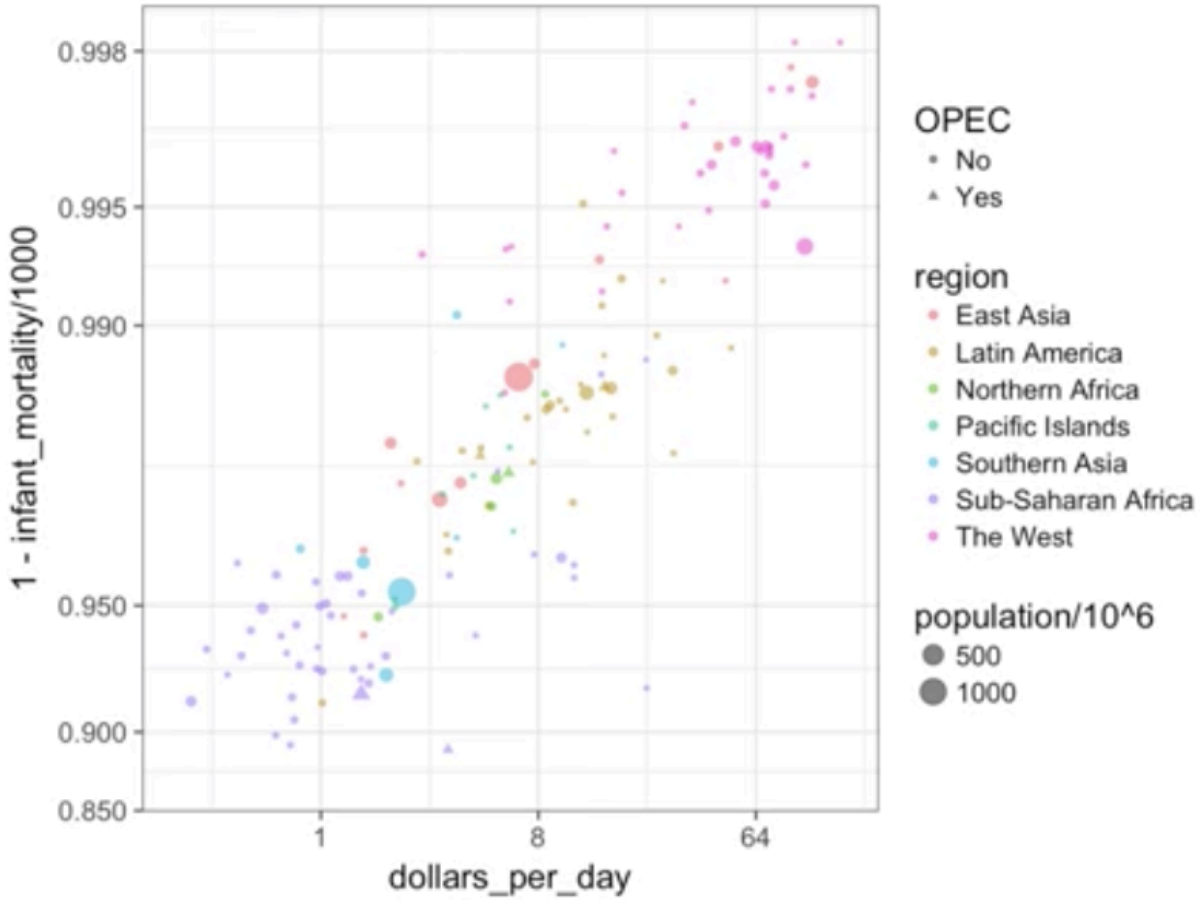
Bland – Altman Plot

Encoding a third variable

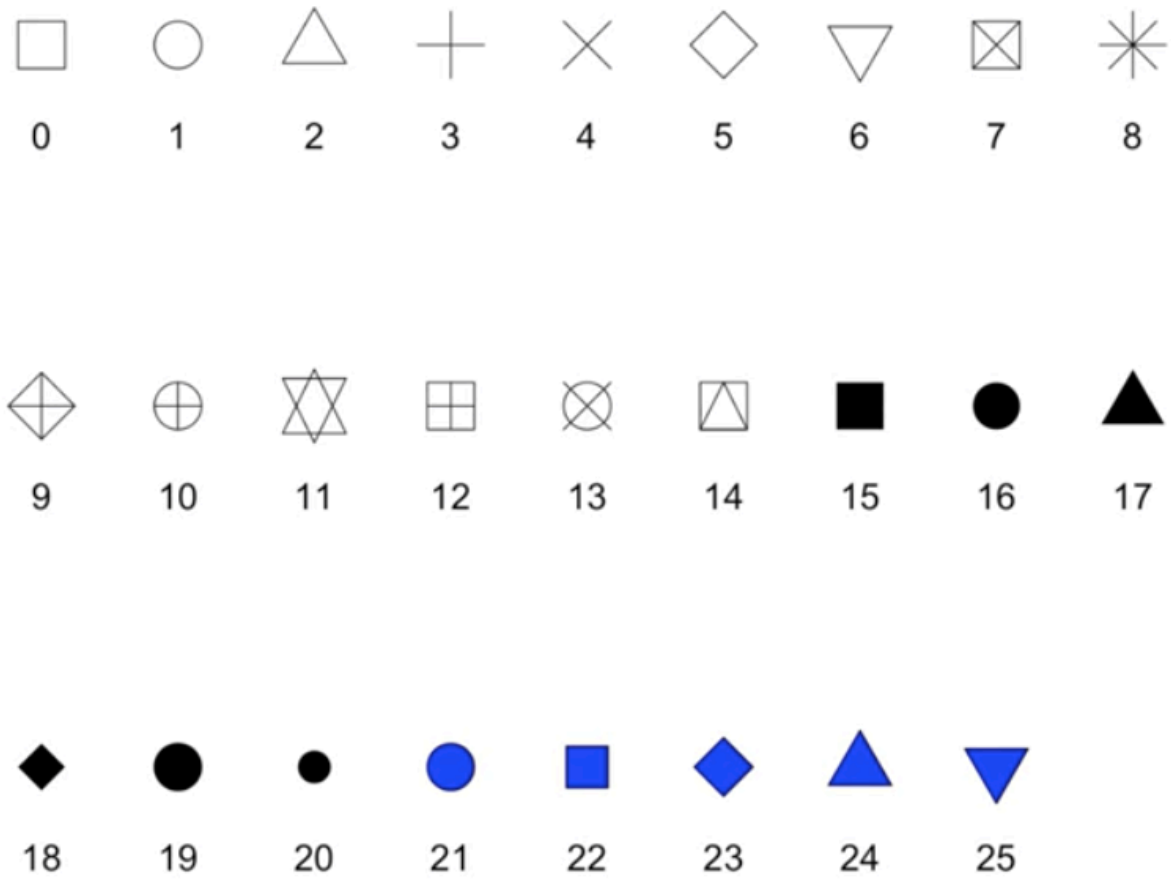
Principle 12

- Include more variables for better understanding

Encoding a third variable

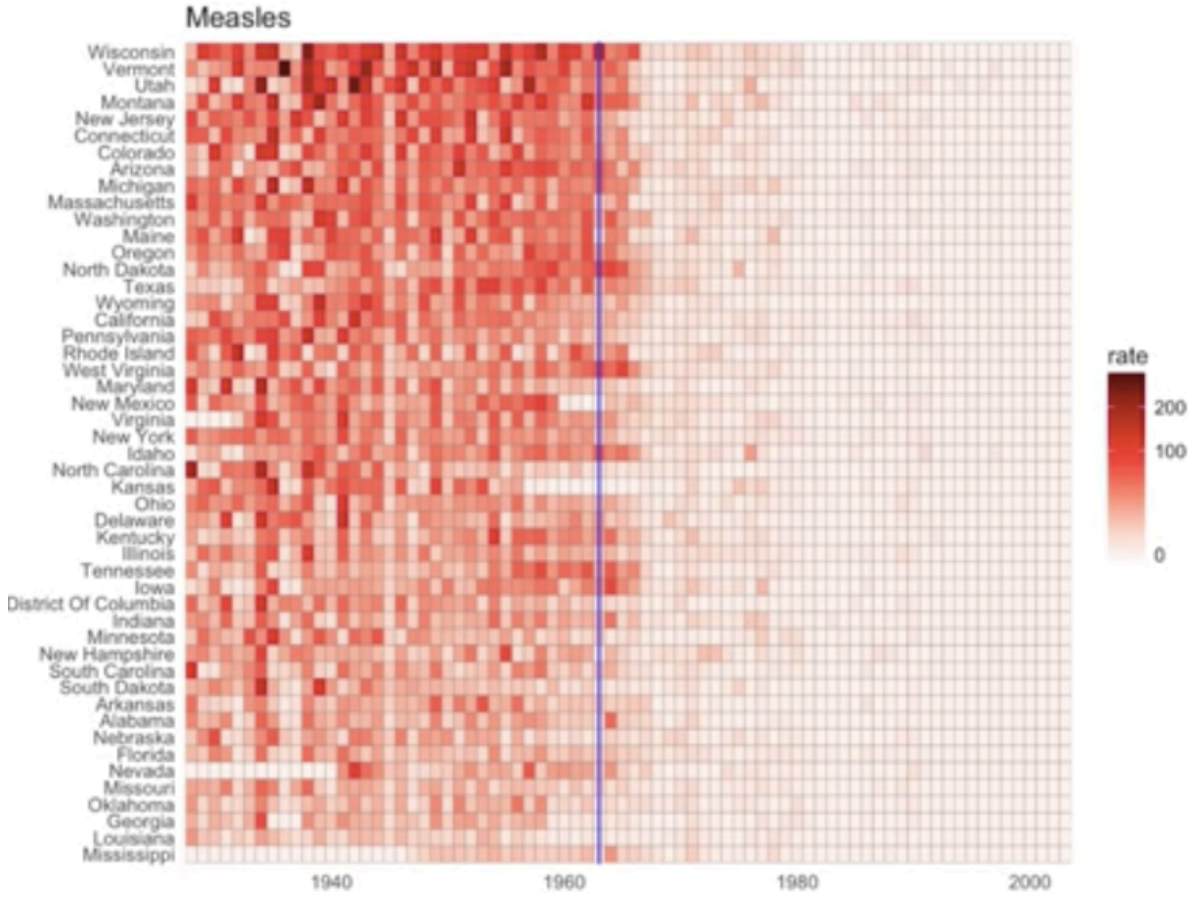


Encoding a third variable



Encoding a third variable

Encoding a third variable (Vaccines data)



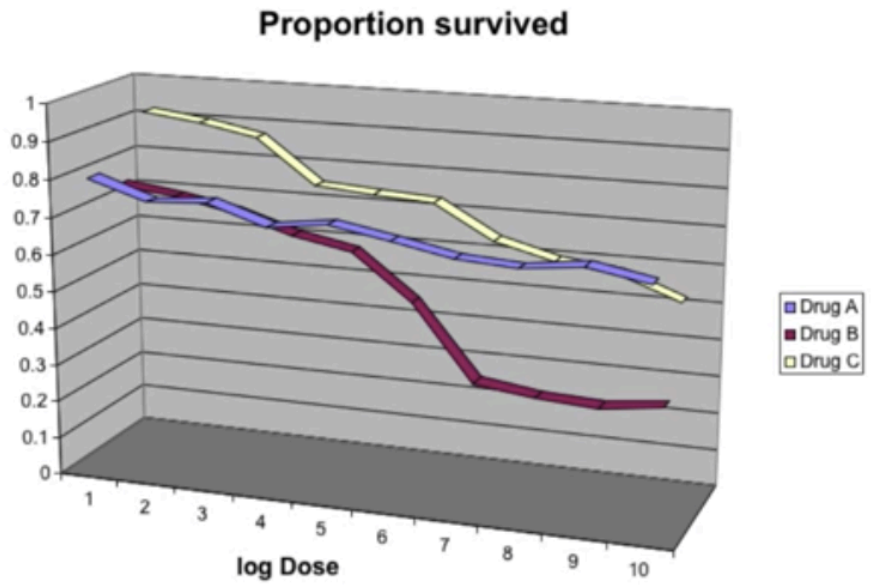
When choosing colors to quantify a numeric variable, we choose between two options, sequential and diverging. Sequential palettes are suited for data that goes from high to low. High values are clearly distinguished from the low values.

On the other hand, diverging colors are used to represent values that verge from a center. We put equal emphasis on both ends of the data range, higher than the center and lower than the center.

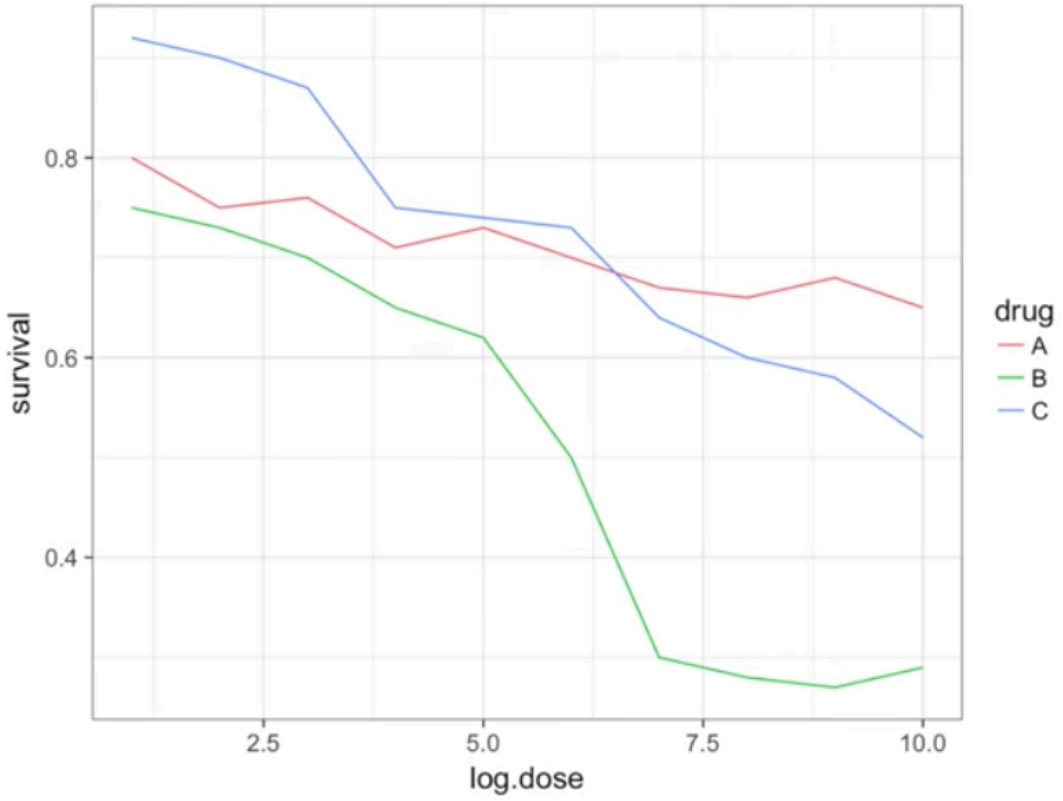
Avoid Pseudo and Gratuitous 3D Plots

Principle 13

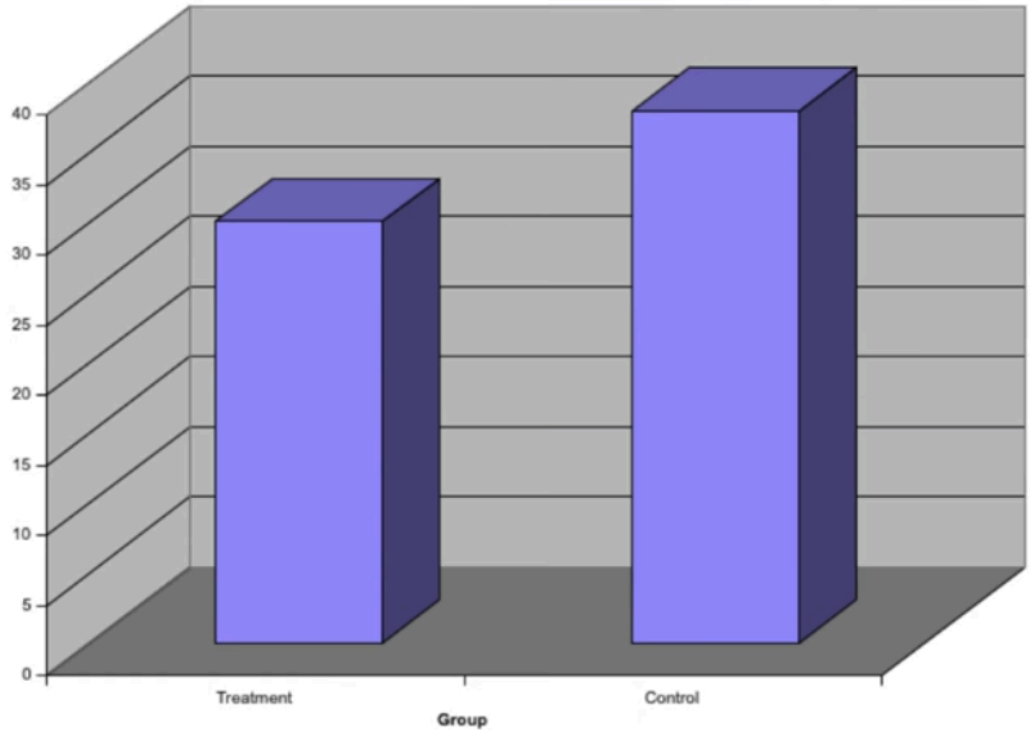
Avoid Pseudo and Gratuitous 3D Plots

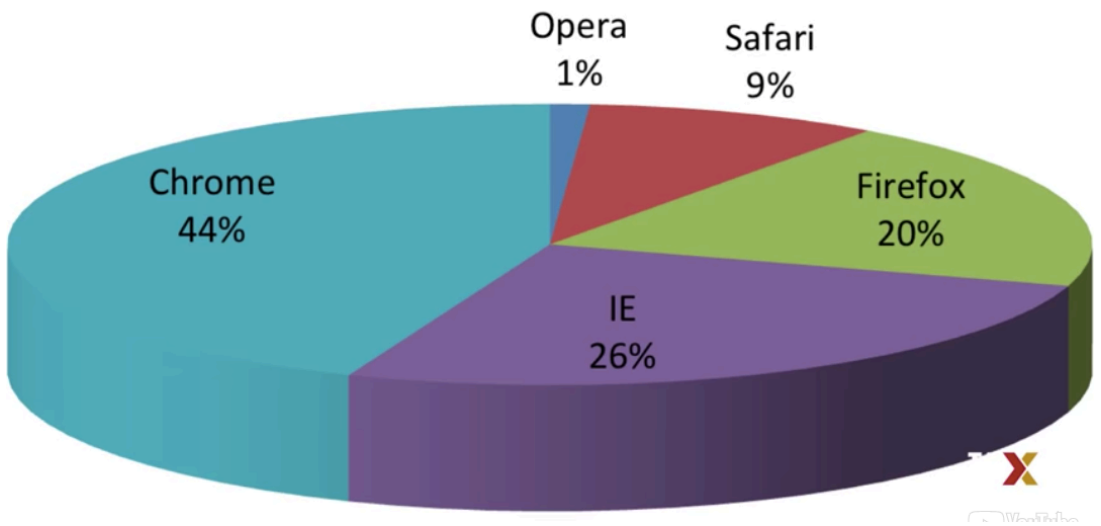


DNA Fingerprinting: A Review of the Controversy Kathryn Roeder Statistical Science Vol. 9, No. 2 (May, 1994), pp. 222-247



Avoid Pseudo and Gratuitous 3D Plots





Avoid Too Many Significant Digits

Principle 14

Avoid Too Many Significant Digits

state	year	Measles	Pertussis	Polio
California	1940	37.8826320	18.3397861	18.3397861
California	1950	13.9124205	4.7467350	4.7467350
California	1960	14.1386471	0.0000000	0.0000000
California	1970	0.9767889	0.0000000	0.0000000
California	1980	0.3743467	0.0515466	0.0515466

state	year	Measles	Pertussis	Polio
California	1940	37.9	18.3	18.3
California	1950	13.9	4.7	4.7
California	1960	14.1	0.0	0.0
California	1970	1.0	0.0	0.0
California	1980	0.4	0.1	0.1

Avoid Too Many Significant Digits

- Place values being compared in columns and not rows

state	disease	1940	1950	1960	1970	1980
California	Measles	37.9	13.9	14.1	1	0.4
California	Pertussis	18.3	4.7	0.0	0	0.1
California	Polio	18.3	4.7	0.0	0	0.1

state	year	Measles	Pertussis	Polio
California	1940	37.9	18.3	18.3
California	1950	13.9	4.7	4.7
California	1960	14.1	0.0	0.0
California	1970	1.0	0.0	0.0
California	1980	0.4	0.1	0.1