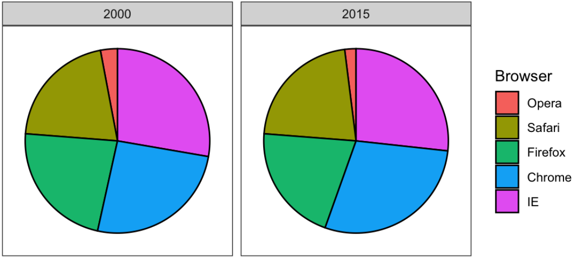
STAT718/BIOL703: Genomic Data Science
Data Visualization Principles

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

Encoding Data Using Visual Cues

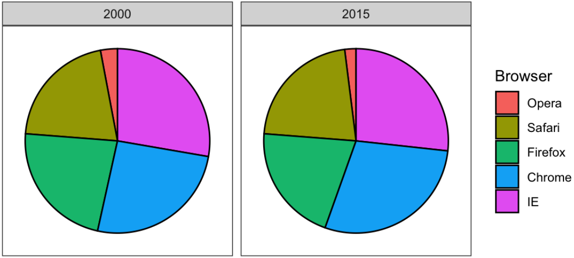
Visual cues: position, aligned lengths, angles, area, brightness, and color hue.



Both area and angle are proportional to the quantity the slice represents.

Encoding Data Using Visual Cues

Visual cues: position, aligned lengths, angles, area, brightness, and color hue.

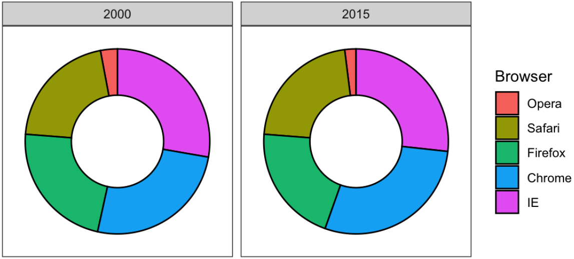


Both area and angle are proportional to the quantity the slice represents.

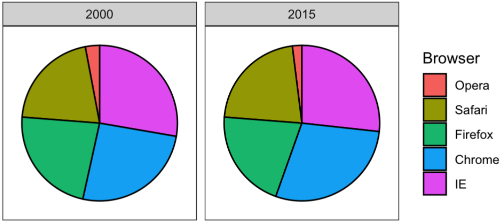
A sub-optimal choice! Humans are not good at precisely quantifying angles and are even worse with area.

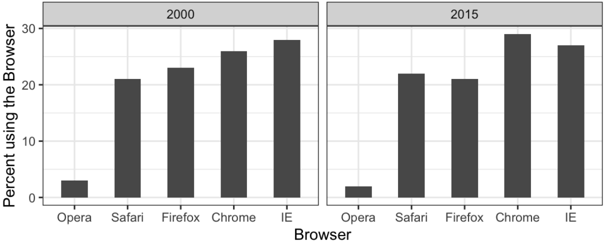
Visual Cues

Can you determine the actual percentages and rank the browsers' popularity?
Can you see how the percentages changed from 2000 to 2015? It is not easy to tell from the plot.



A better choice





It is easier to see the differences in the barplot.

Visual Cues

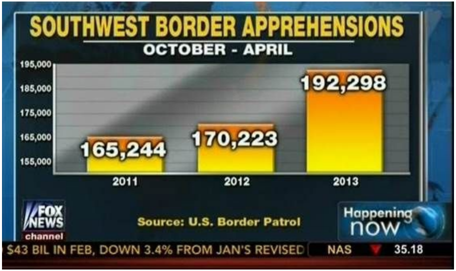
Visual cues: position, aligned lengths, angles, area, brightness, and color hue.

In general, when displaying quantities, position and length are preferred over angles and/or area.

Brightness and color are even harder to quantify than angles. But they can be useful when two dimensions are displayed together.

Know when to include zero

By avoiding 0, relatively small differences can be made to look much bigger than they actually are.

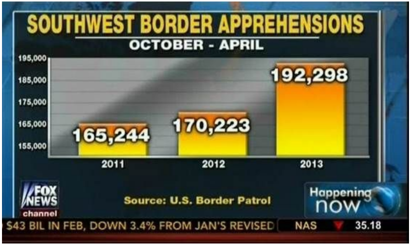


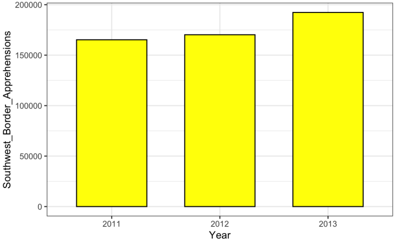
Source: Fox News, via Media Matters[1]

[1] http://mediamatters.org/blog/2013/04/05/fox-news-newest-dishonest-chart-immigration-enf/193507

Know when to include zero

In reality, it only increased by about ≈ 16%. Starting the graph at 0 illustrate this clearly.



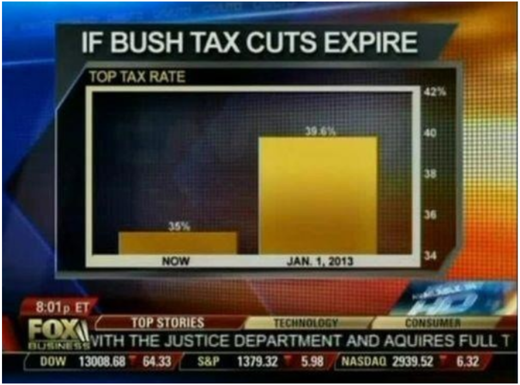


source: Fox News, via Media Matters[2]

[2] http://mediamatters.org/blog/2013/04/05/fox-news-newest-dishonest-chart-immigration-enf/193507

Another Example

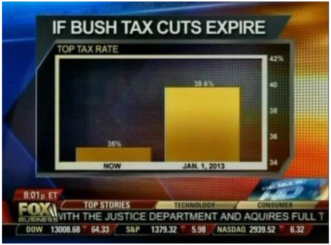
This plot makes a 13% increase look like a five fold change.

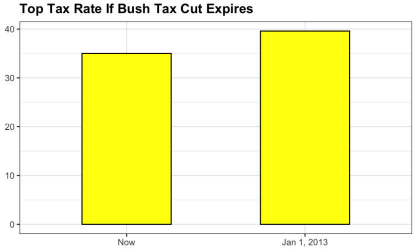


source: Fox News, via Flowing Data[3]

[3] http://flowingdata.com/2012/08/06/fox-news-continues-charting-excellence

Know when to include zero



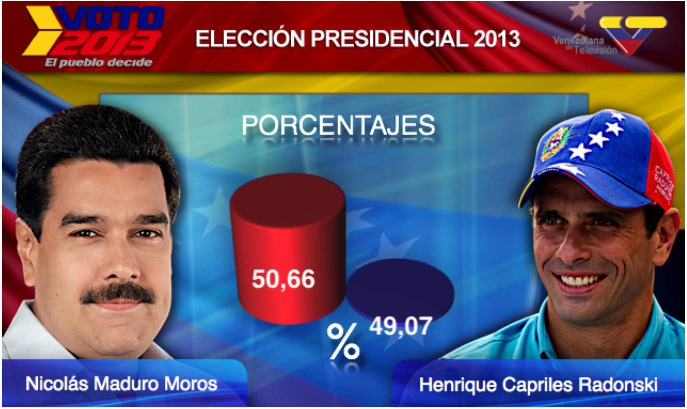


source: Fox News, via Flowing Data[4]

[4] http://flowingdata.com/2012/08/06/fox-news-continues-charting-excellence

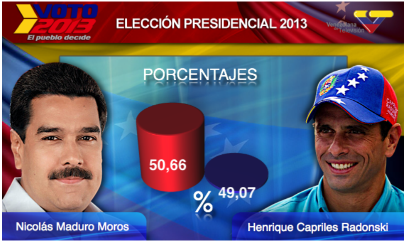
Extreme Example

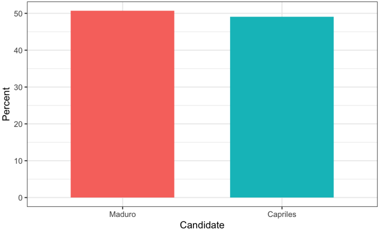
An extremely example that make an under 2% difference look like a 10-100 fold change.



Source: Venezolana de Televisión via El Mundo9.

Appropriate plot





Know when to include zero

When using position rather than length, it is then not necessary to include 0.
Comparing differences between groups relative to the within-group variability.

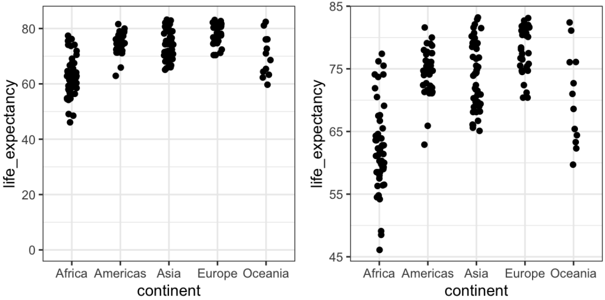
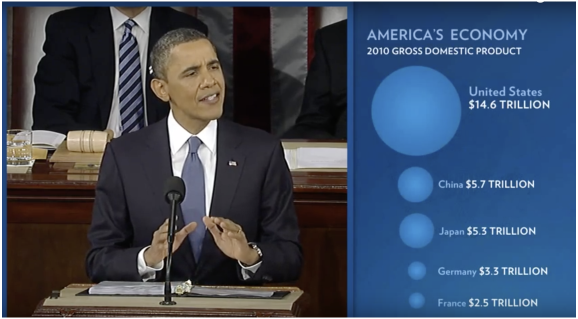


Figure 1: country average life expectancy stratified across continents in 2012

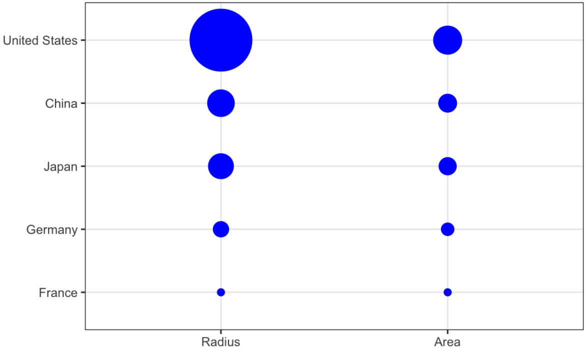
Do not distort quantities

Proportional to area not radius.



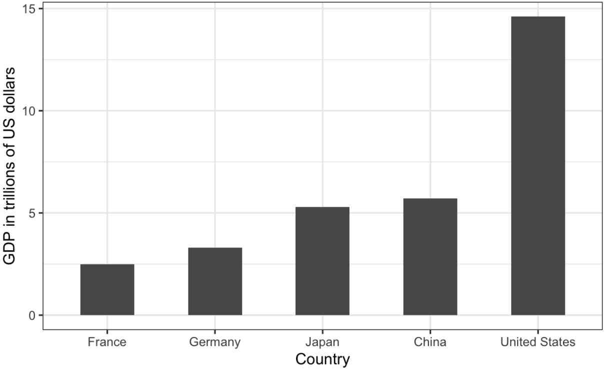
Judging by the area of the circles, the US appears to have an economy over five times larger than China's and over 30 times larger than France's.
source:The 2011 State of the Union Address[5]

[5] https://www.youtube.com/watch?v=kl2g40GoRxg

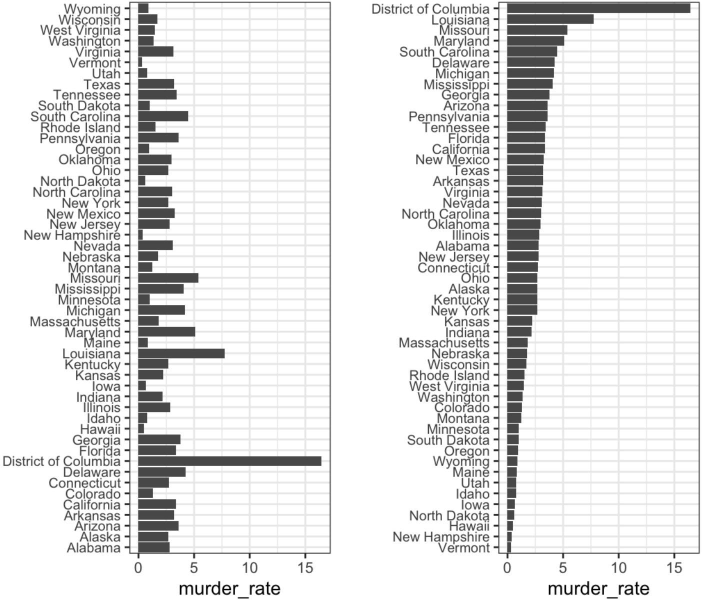

United States
China
Japan
Germany
France
Radius
Area

Better plot

Use position and length.

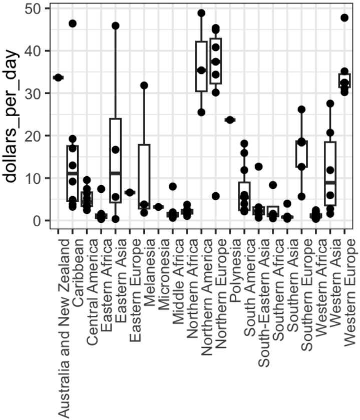


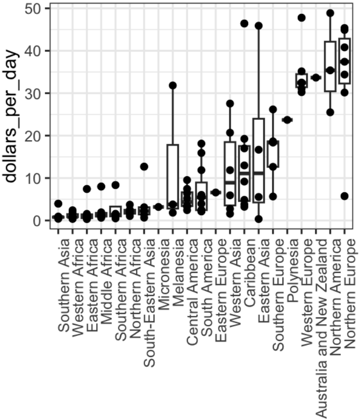
Order categories by a meaningful value



Income distributions across regions

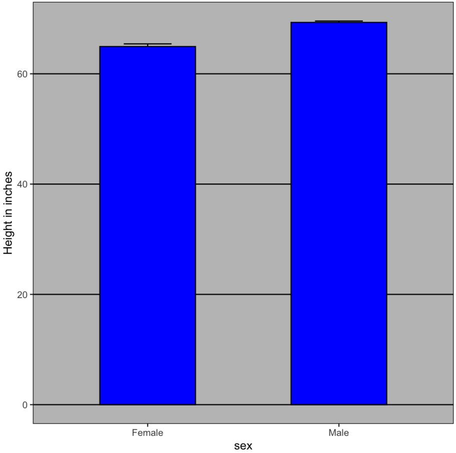
The first orders the regions alphabetically, while the second orders them by the group's median.



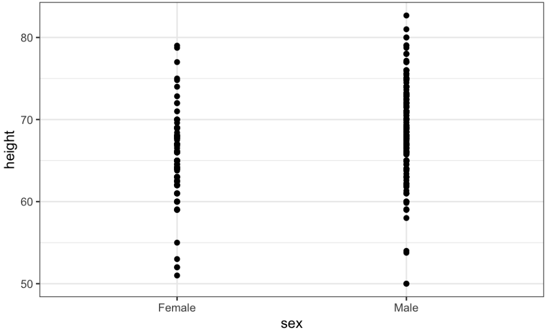


Show the Data

The dynamite plot

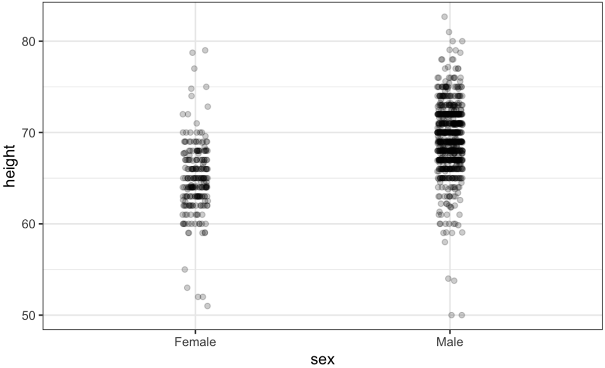


Show The Data: A Second Plot

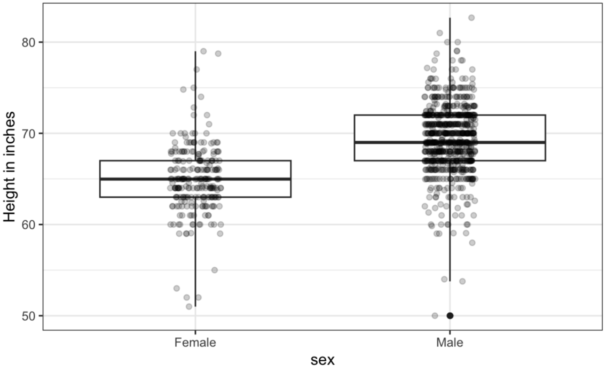


Show The Data

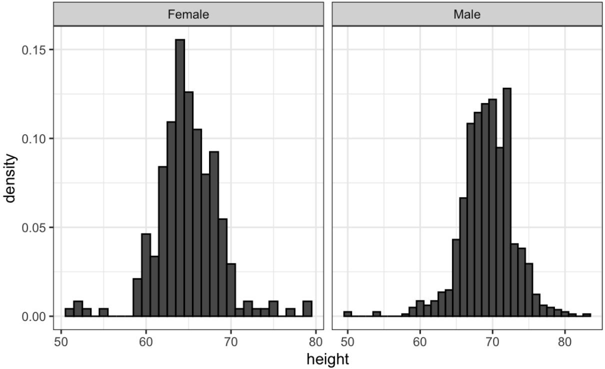
- Jittter
- Alpha blending



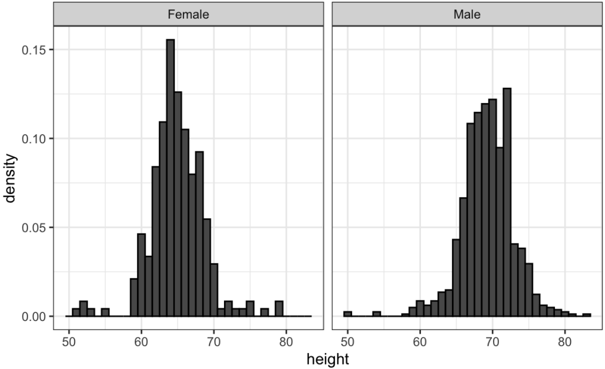
Show The Data: Jitter with boxplot



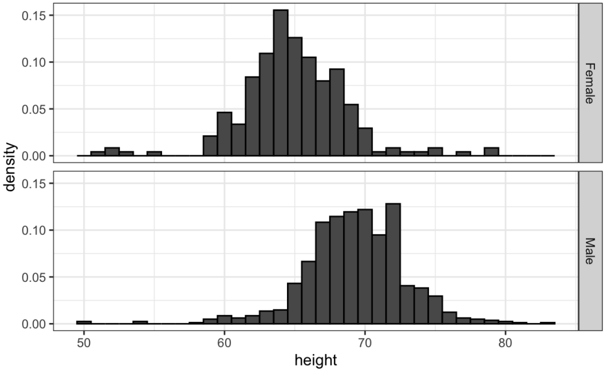
Show The Data: Historgram



Show The Data: Common Axes

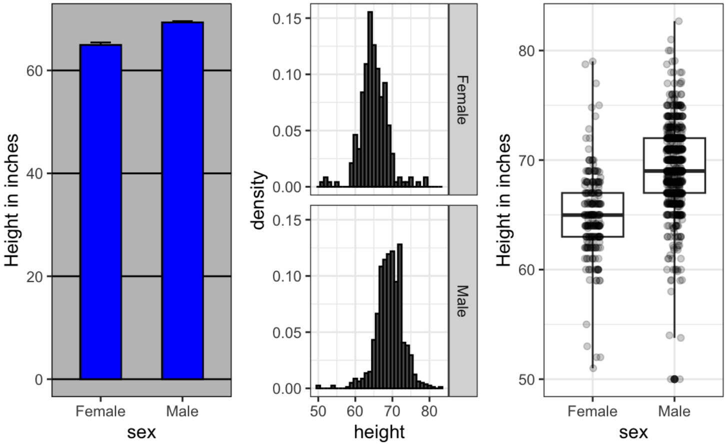


Show The Data: Aligning the plots



Ease Comparison

- Use common axes
- Aligning plots



Transformation

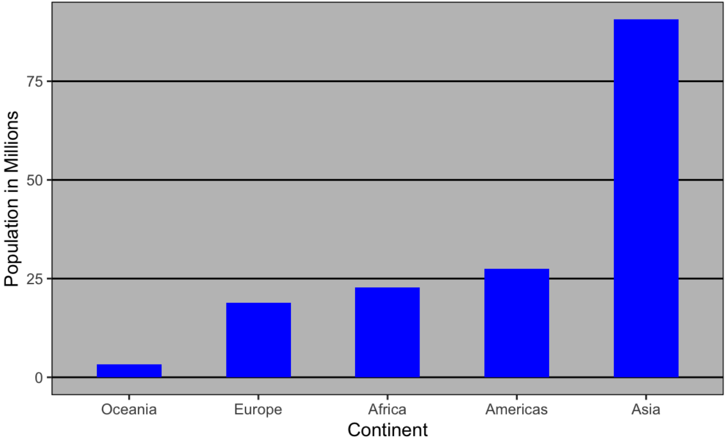


Figure 2: Average population sizes in 2015.

Transformation

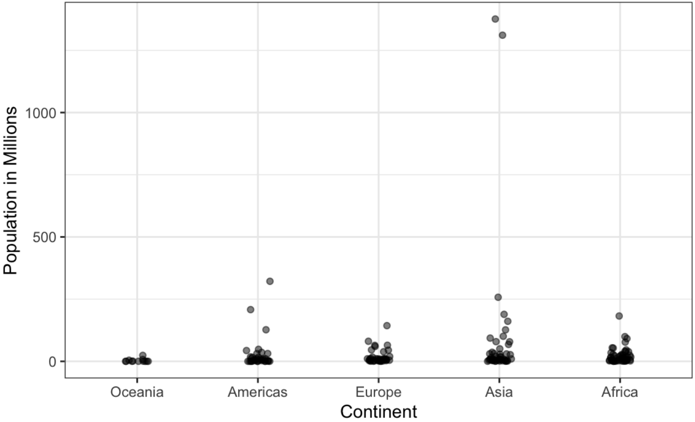


Figure 3: Average population sizes in 2015.

log Transformation

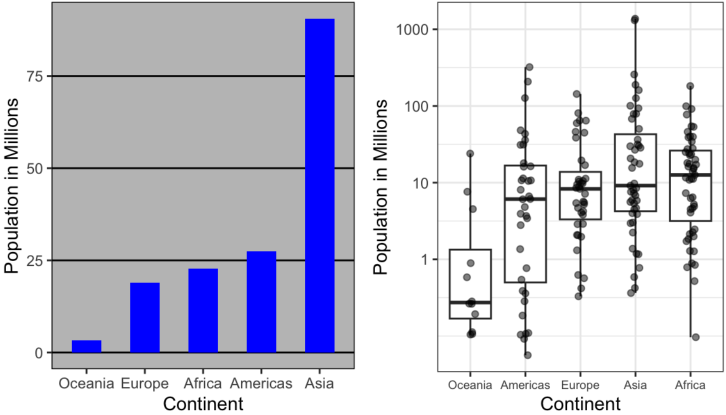


Figure 4: Average population sizes in log scale in 2015.

With the new plot, we realize that countries in Africa actually have a larger median population size than those in Asia.

Visual cues to be compared should be adjacent

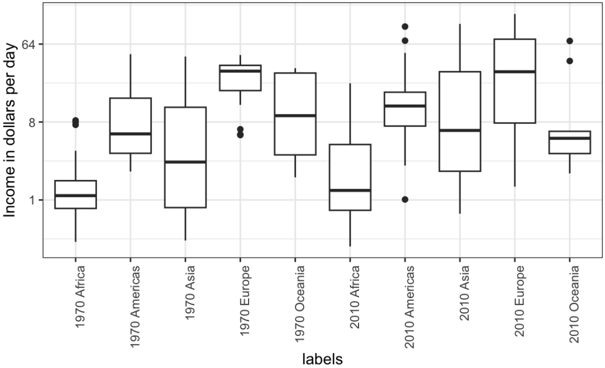


Figure 5: Labels order alphabetically

Visual cues to be compared should be adjacent

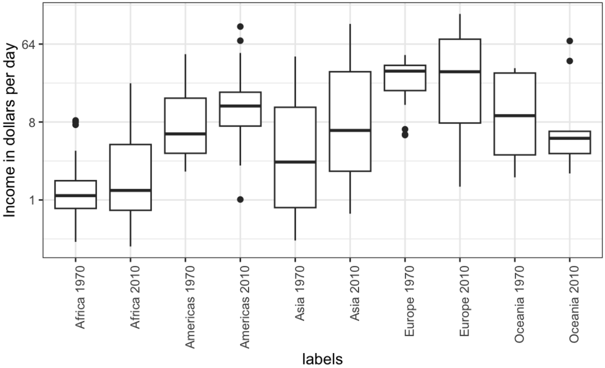


Figure 6: 1970 vs 2010

Visual cues to be compared should be adjacent

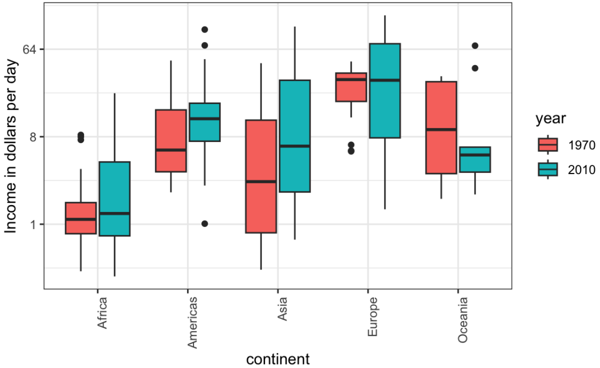


Figure 7: 1970 vs 2010 with color

Think of the color blind

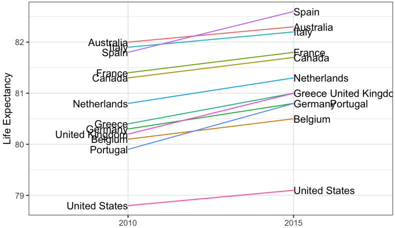
```
>color_blind_friendly_cols <-
  c("#999999", "#E69F00", "#56B4E9", "#009E73",
    "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
```

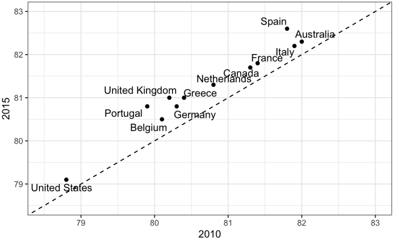
Plots for two variables

- scatterplots
- slope chart
- Bland-Altman plot

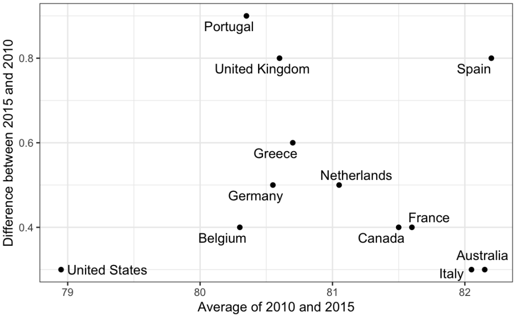
Slope Charts

An advantage of the slope chart is that it permits us to quickly get an idea of changes based on the slope of the lines.

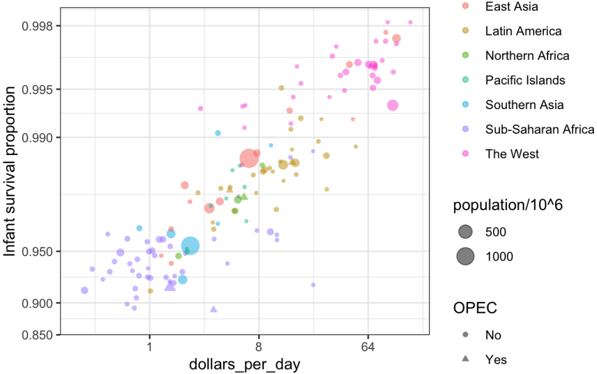




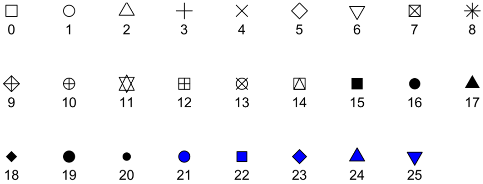
Bland-Altman plot



Encoding A Third Variable



Encoding Categorical Variables: Shape and Color



Continuous Variables: Color, Intensity, size

- Sequential: values from high to low
- Diverging: diverge from a center

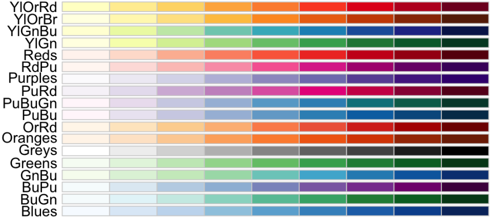


Figure 8: Sequential colors

Divergent Color Patterns

Equal emphasis on both ends of the data range.

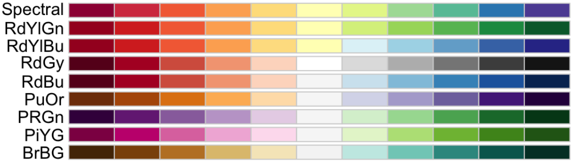
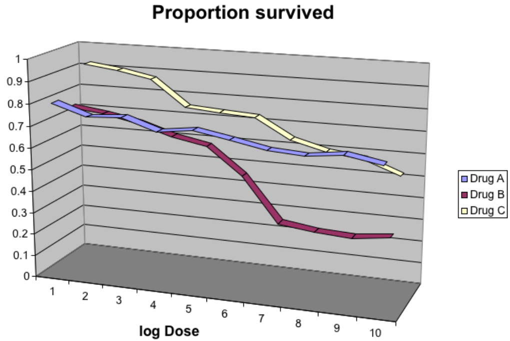


Figure 9: Divergent patterns

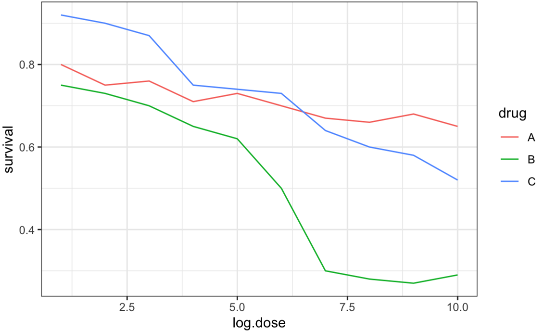
Avoid pseudo-three-dimensional plots

Can you tell when the purple ribbon intersects the red one?

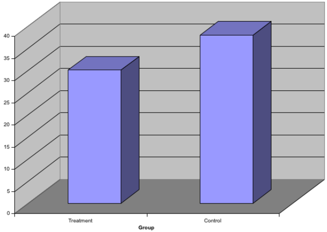


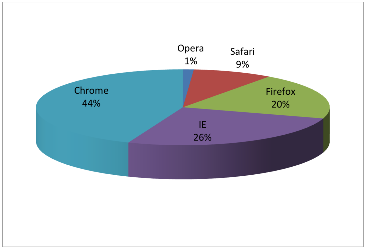
Source:
https://projecteuclid.org/download/pdf_1/euclid.ss/1177010488

Use Color to represent the categorical variable

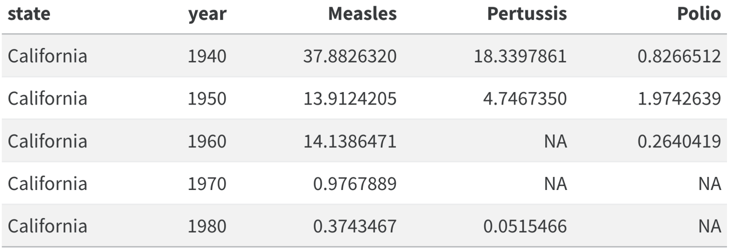


The 3rd dimension does not represent a quantity and only adds confusion.





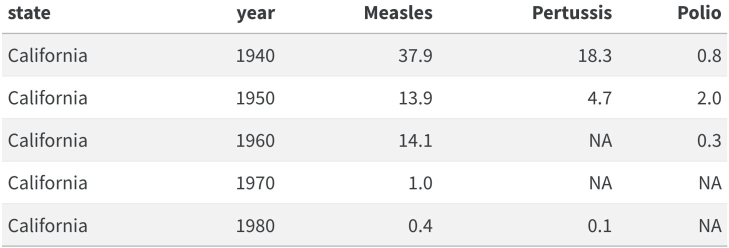
Avoid too many significant digits



state	year	Measles	Pertussis	Polio
California	1940	37.8826320	18.3397861	0.8266512
California	1950	13.9124205	4.7467350	1.9742639
California	1960	14.1386471	NA	0.2640419
California	1970	0.9767889	NA	NA
California	1980	0.3743467	0.0515466	NA

Figure 10: Per 10,000 disease rates

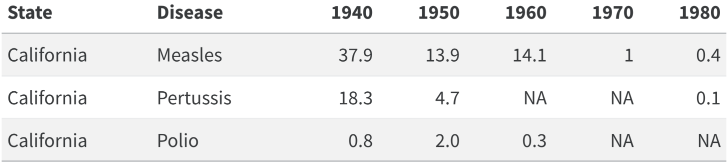
Avoid too many significant digits



state	year	Measles	Pertussis	Polio
California	1940	37.9	18.3	0.8
California	1950	13.9	4.7	2.0
California	1960	14.1	NA	0.3
California	1970	1.0	NA	NA
California	1980	0.4	0.1	NA

Figure 11: Per 10,000 disease rates

Place values being compared on columns rather than 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	NA	NA	0.1
California	Polio	0.8	2.0	0.3	NA	NA

Figure 12: Per 10,000 disease rates