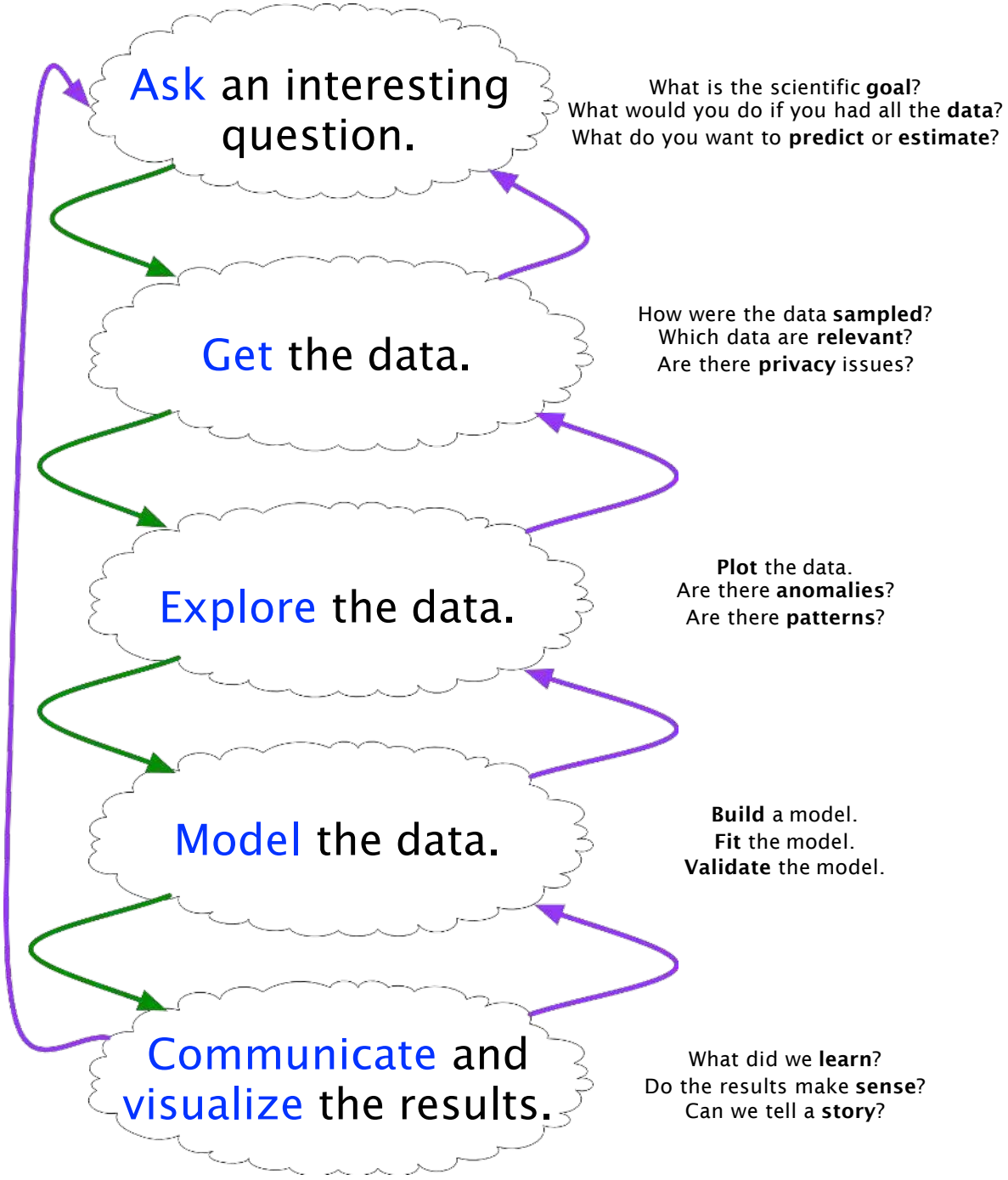
# Data Exploration and Visualization

# Data Exploration

Not always sure what we are looking for (until we find it)

# Visualization

- Visualization is the conversion of data into a visual or tabular format.
- Visualization helps understand the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns


Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?
Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?
Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

# What is Data?

- Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

**Attributes**

**Objects**

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Type of variables (attributes)

- **Descriptive (categorical) variables**
  - **Nominal** variables (no order between values): gender, eye color, race group, ...
  - **Ordinal** variables (inherent order among values): response to treatment: none, slow, moderate, fast
- **Measurement variables**
  - **Continuous measurement** variable: height, weight, blood pressure ...
  - **Discrete measurement** variable (values are integers): number of siblings, the number of times a person has been admitted to a hospital ...

# Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

# 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 |

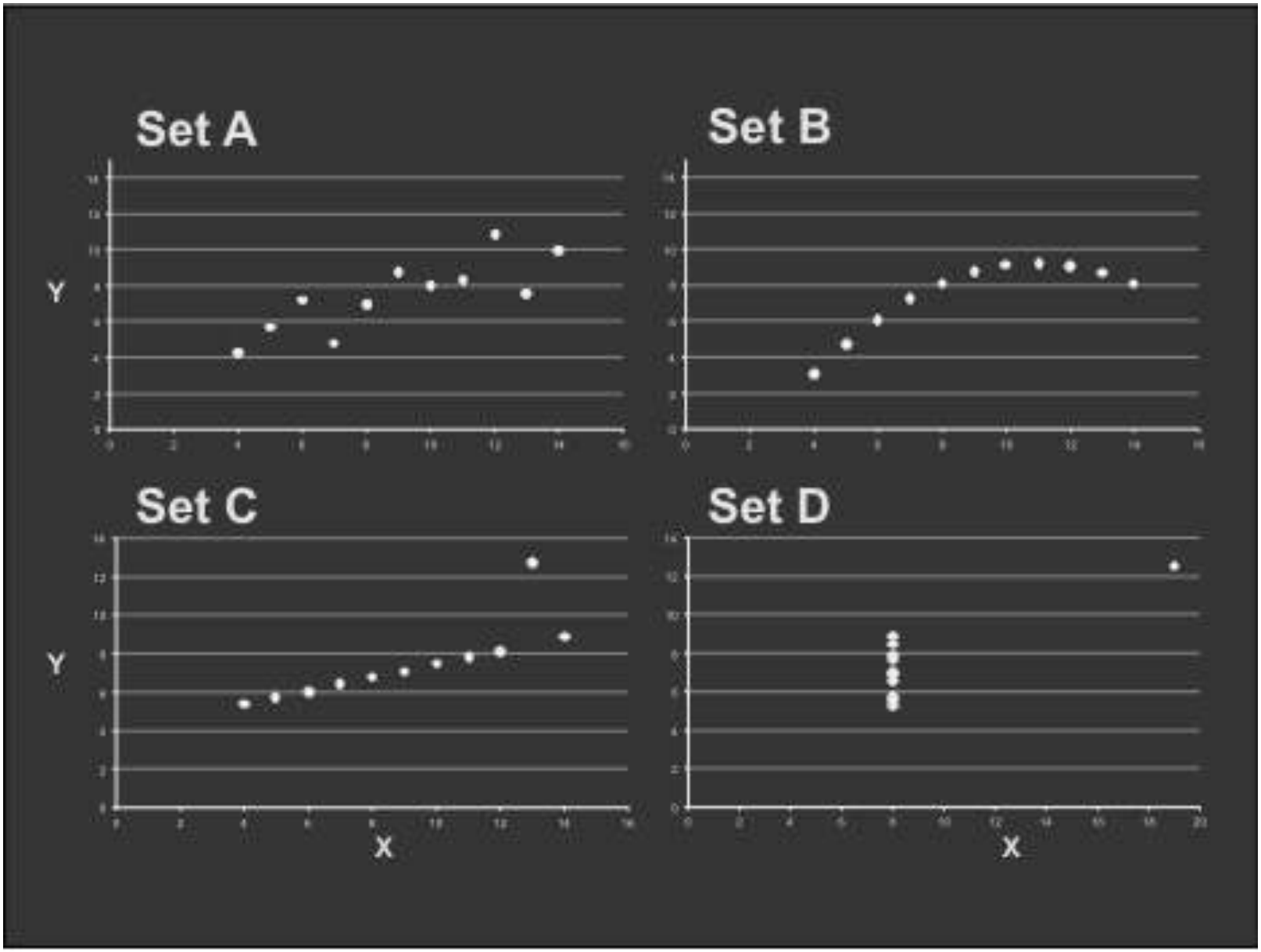
**Summary Statistics Linear Regression**

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\ X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

[Anscombe 73]

# Looking at Data



# Chart types

- Single variable
  - Dot plot
  - Box-and-whisker plot
  - Histogram
  - Jitter plot
  - Error bar plot
  - Cumulative distribution function

# Chart types

- Two variables
  - Bar chart
  - Scatter plot
  - Line plot
  - Log-log plot
- More than two variables
  - Stacked plots
  - Parallel coordinate plot

# Sample Data

| Height | Weight | Waist | Hip | bp.sys | bp.dia |
|---|---|---|---|---|---|
| 172 | 72 | 87 | 94 | 127.5 | 80 |
| 166 | 91 | 109 | 107 | 172.5 | 100 |
| 174 | 80 | 95 | 101 | 123 | 64 |
| 176 | 79 | 93 | 100 | 117 | 76 |
| 166 | 55 | 70 | 94 | 100 | 60 |
| 163 | 76 | 96 | 99 | 160 | 87.5 |
| 154 | 84 | 98 | 118 | 130 | 80 |
| 165 | 90 | 108 | 101 | 139 | 80 |
| 155 | 66 | 80 | 96 | 120 | 70 |
| 146 | 59 | 77 | 96 | 112.5 | 75 |
| 164 | 62 | 76 | 93 | 130 | 47.5 |
| 159 | 59 | 76 | 96 | 109 | 69 |
| 163 | 69 | 96 | 99 | 155 | 100 |
| 143 | 73 | 97 | 117 | 137.5 | 85 |
| . . . | | | | | |

# Plotting a Vector

- *plot(v)* will print the elements of the vector v according to their index

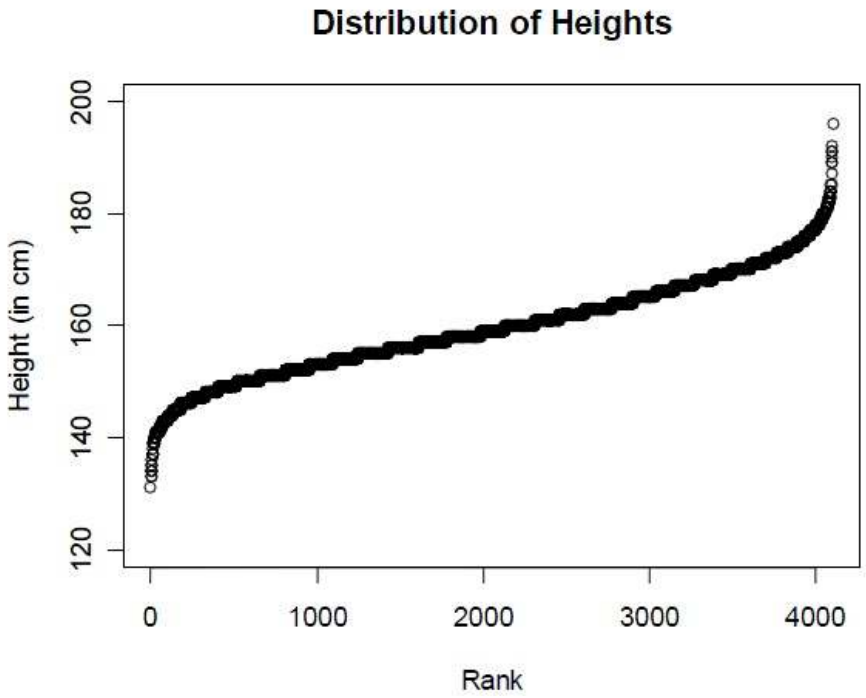
```
# Plot height for each observation
> plot(dataset$Height)

# Plot values against their ranks
> plot(sort(dataset$Height))
```

# Parameters for plot()

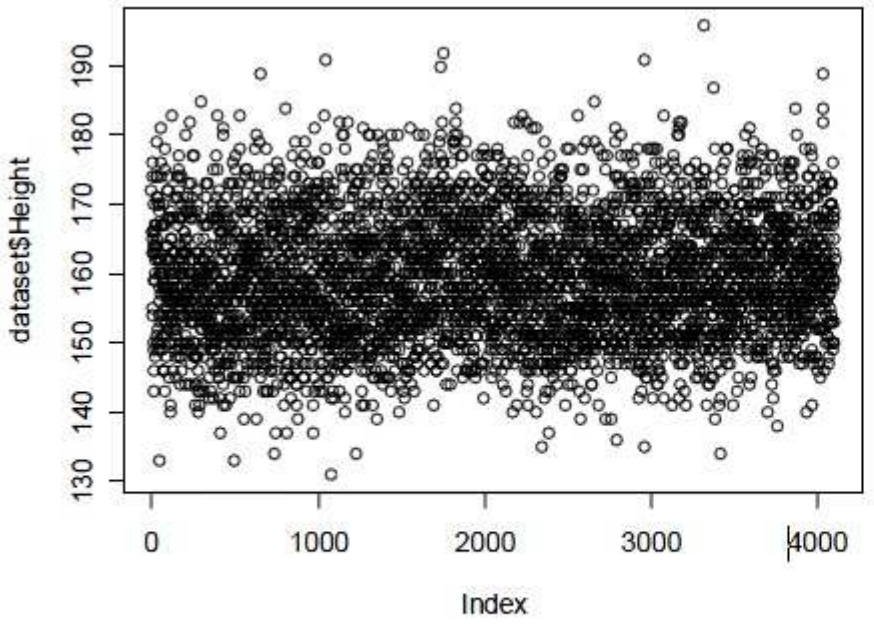
- Specifying labels:
  - main – provides a title
  - xlab – label for the x axis
  - ylab – label for the y axis

- Specifying range limits
  - ylim – 2-element vector gives range for x axis
  - xlim – 2-element vector gives range for y axis

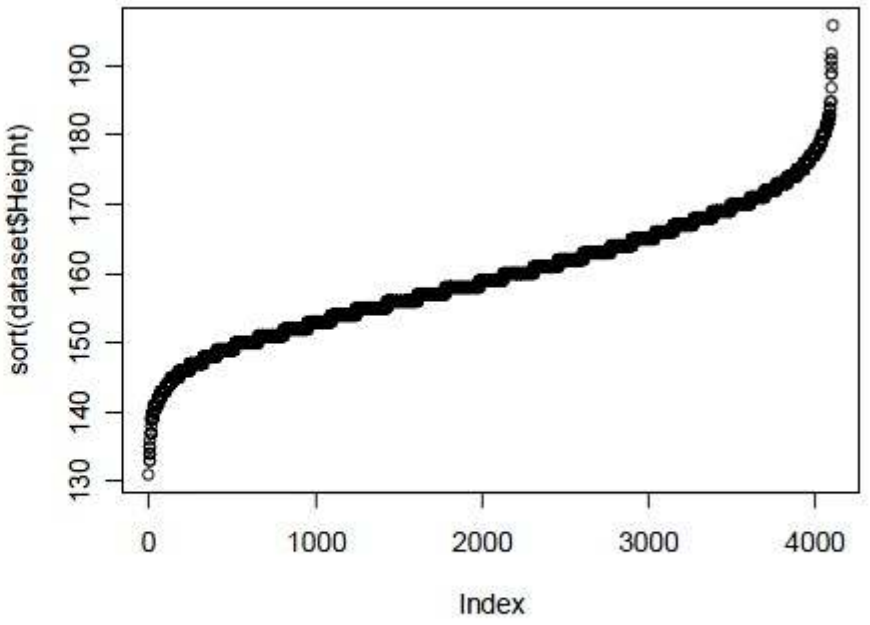



```
plot(sort(dataset$Height), ylim = c(120,200),
ylab = "Height (in cm)", xlab = "Rank", main = "Distribution of Heights")
```

# Plotting a Vector



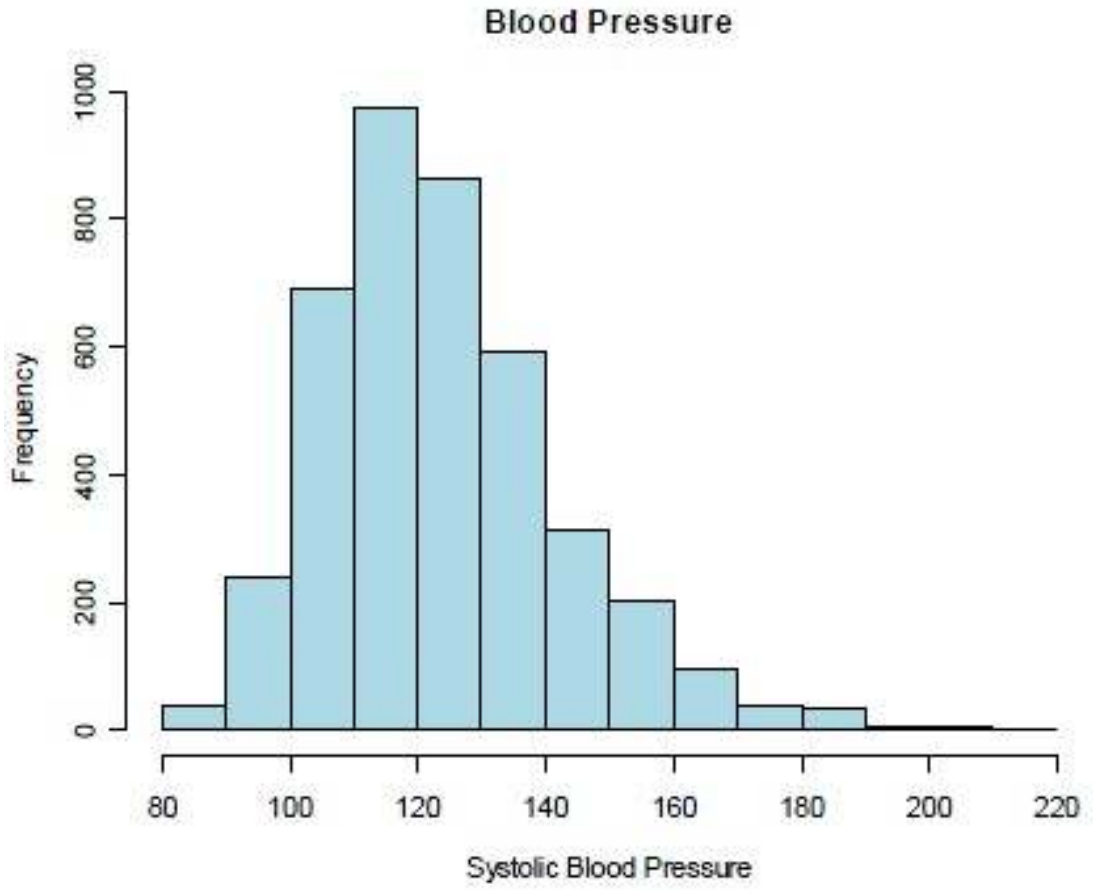

```
plot(dataset$Height)
```




```
plot(sort(dataset$Height))
```

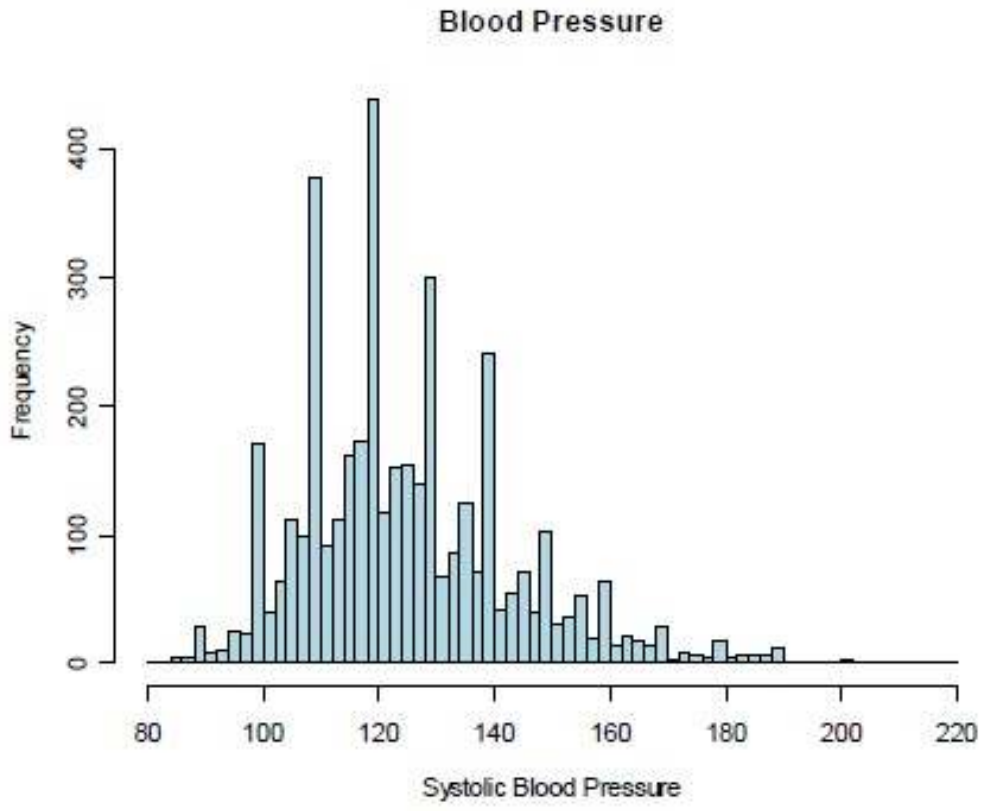
# Histogram

```
hist(dataset$bp.sys, col = "lightblue",
xlab = "Systolic Blood Pressure", main = "Blood Pressure")
```



# Histogram

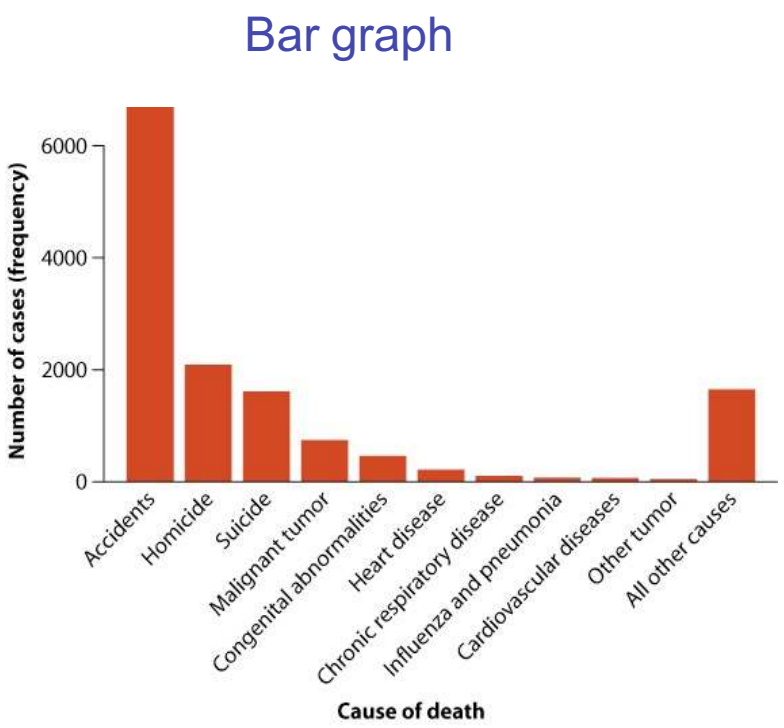
```
hist(dataset$bp.sys, col = "lightblue", breaks = seq(80,220,by=2),
     xlab = "Systolic Blood Pressure", main = "Blood Pressure")
```



# Bar graph

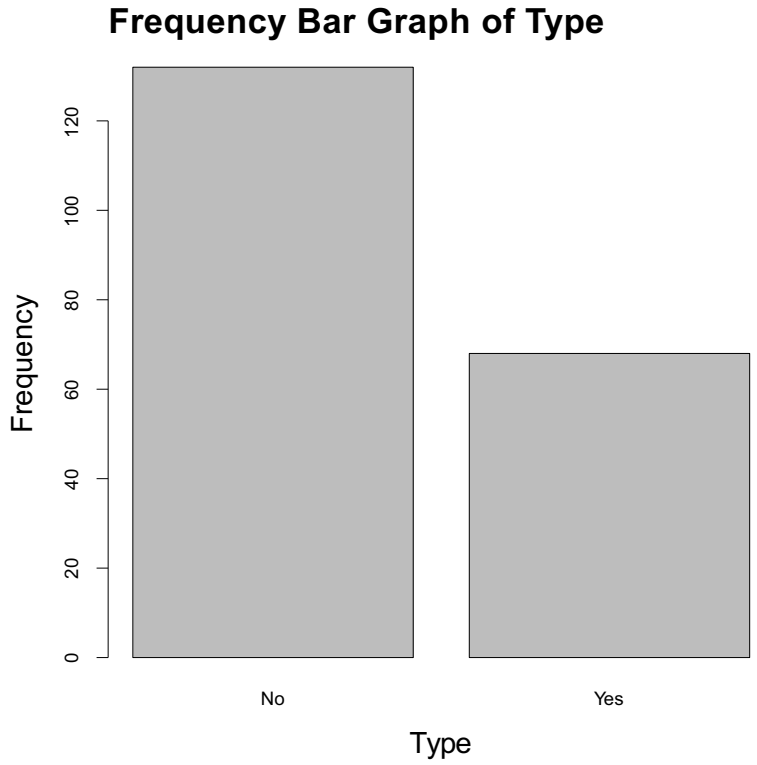
| Cause of death | Frequency |
|---|---|
| Accidents | 6,688 |
| Homicide | 2,093 |
| Suicide | 1,615 |
| Malignant tumor | 745 |
| Heart disease | 463 |
| Congenital abnormalities | 222 |
| Chronic respiratory disease | 107 |
| Influenza and pneumonia | 73 |
| Cerebrovascular diseases | 67 |
| Other tumor | 52 |
| All other causes | 1,653 |

Frequency table showing the ten most common causes of death in Americans between 15 and 19 years of age in 1999. The total number of deaths is *n* = 13,778.



# Bar graphs and frequencies

```
> type.freq <- table(Pima.tr$type)
> barplot(type.freq, xlab = "Type", ylab = "Frequency", main = "Frequency Bar Graph of Type")
```

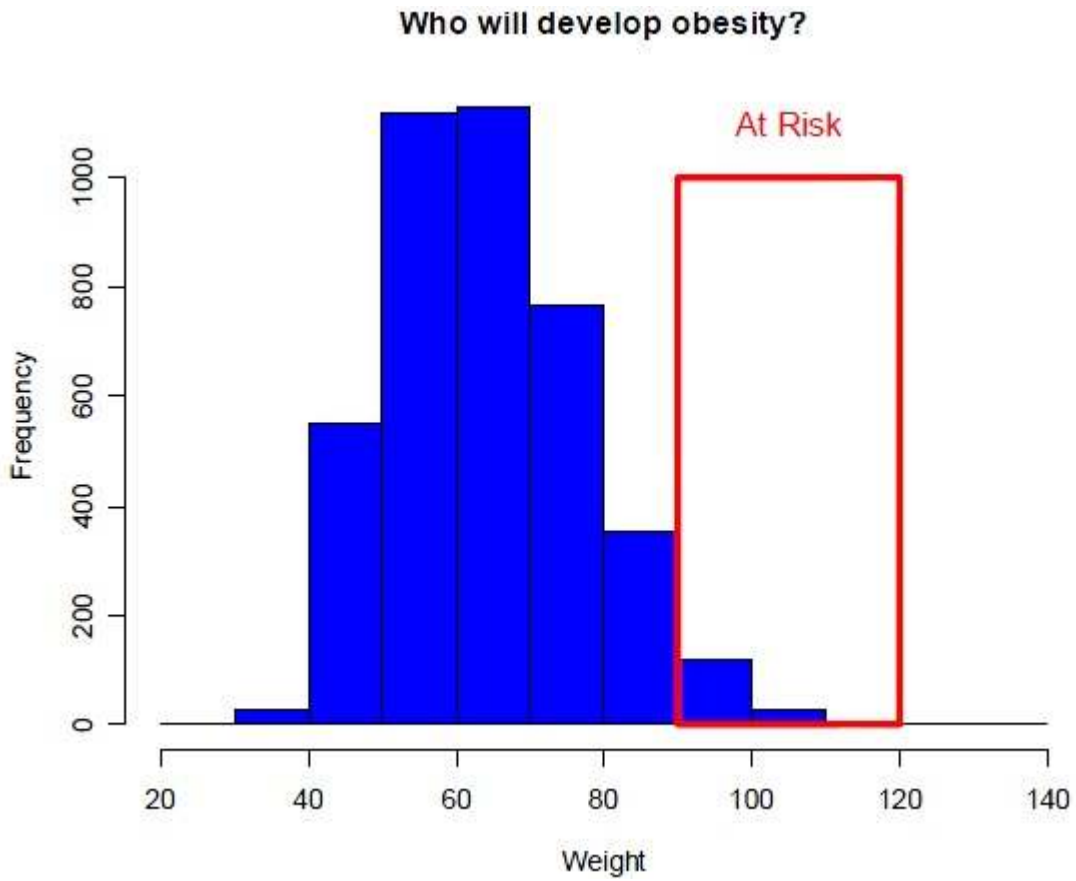


# Adding a Label Inside a Plot

> hist(dataset$Weight, xlab = "Weight", main = "Who will develop obesity?", col = "blue")

> rect(90, 0, 120, 1000, border = "red", lwd = 4)

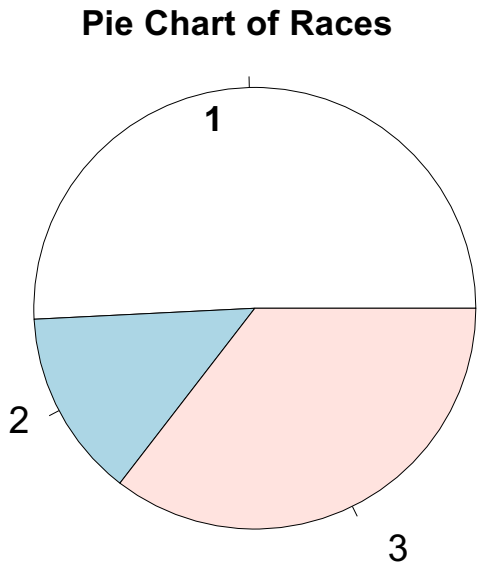
> text(105, 1100, "At Risk", col = "red", cex = 1.25)



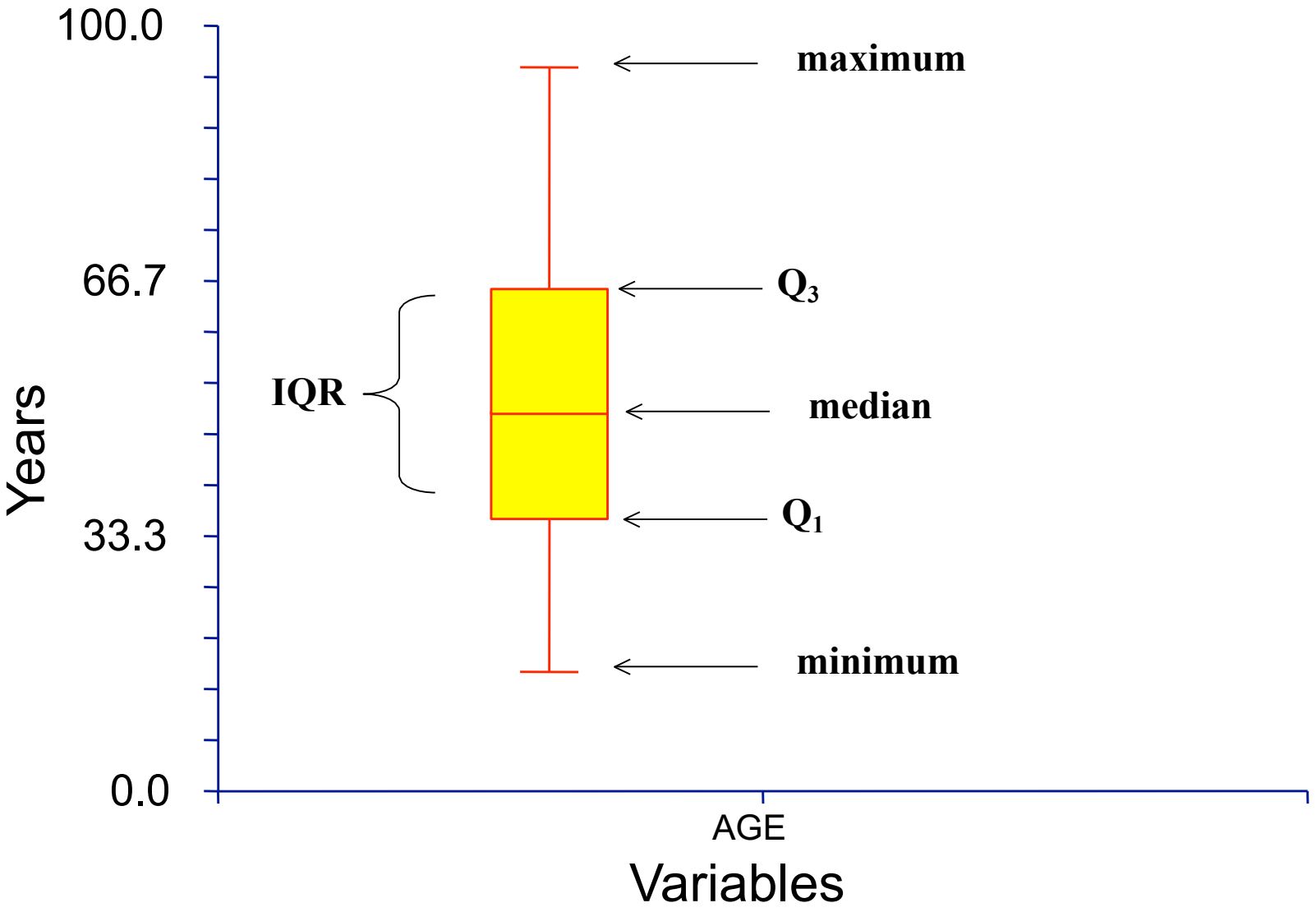
# Pie chart

- We can use a `pie chart` to visualize the relative frequencies of different categories for a categorical variable.
- In a pie chart, the area of a circle is divided into sectors, each representing one of the possible categories of the variable.
- The area of each sector *c* is proportional to its frequency.

```
slices <- c(11, 4,  6)
lbls <- c("1", "2", "3",)
pie(slices, labels = lbls, main="Pie Chart of Races")
```

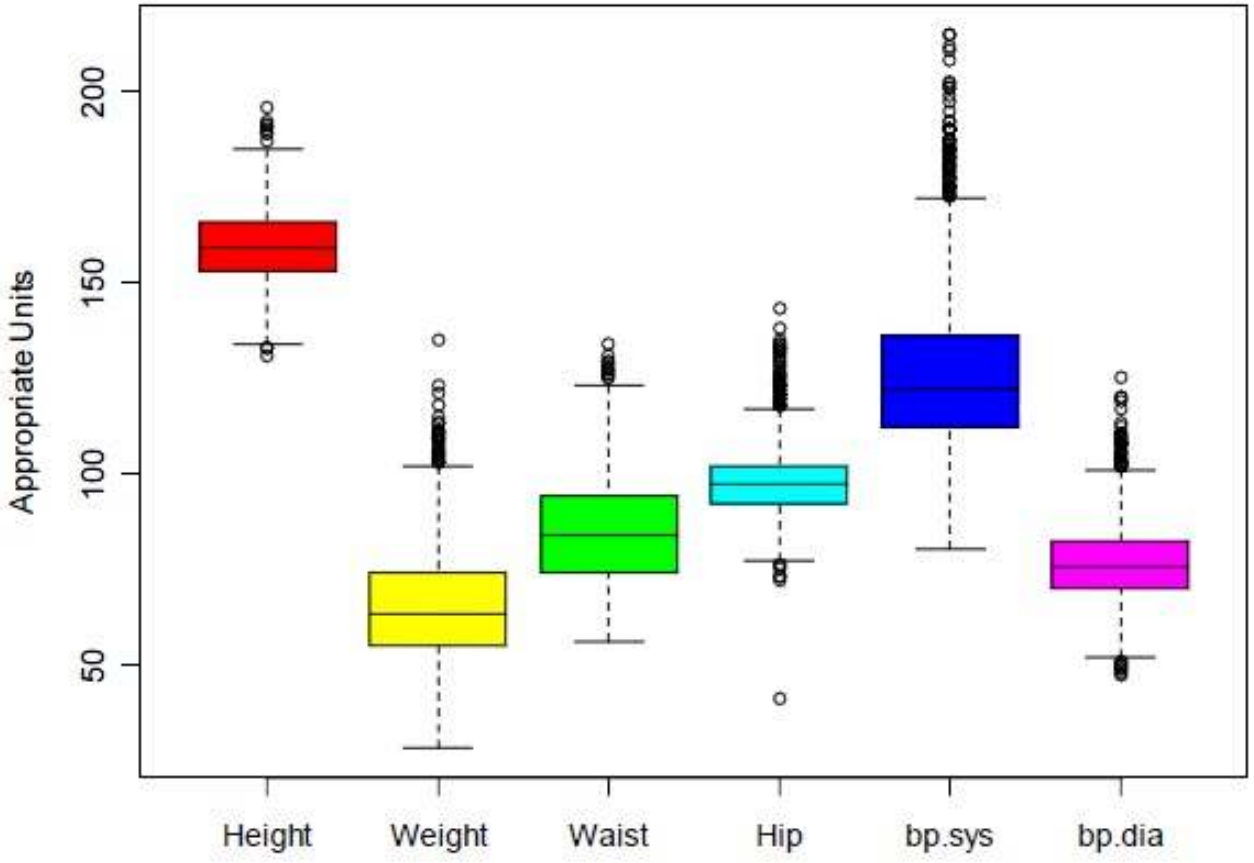


# Box Plots


100.0
66.7
33.3
0.0
Years
maximum
$Q_3$
IQR
median
$Q_1$
minimum
AGE
Variables

# Example of Box Plots

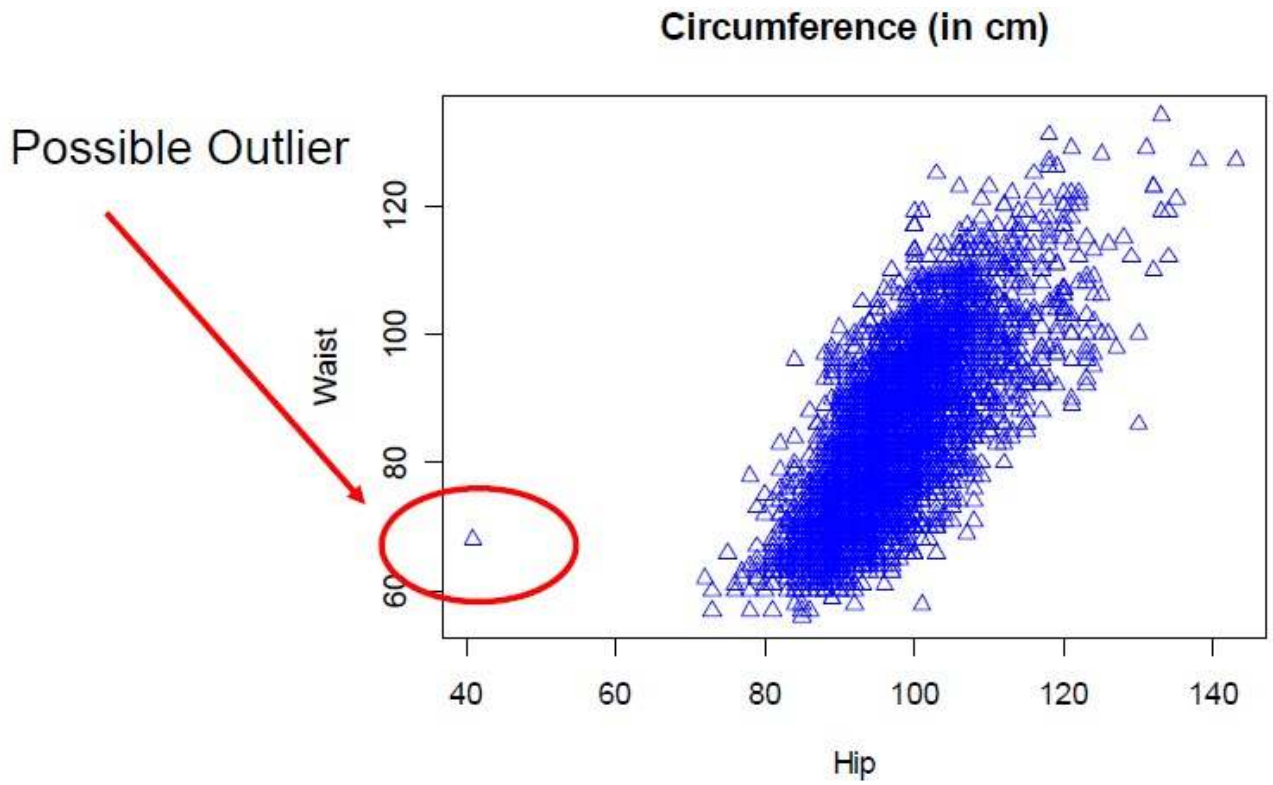
- Box plots can be used to compare attributes



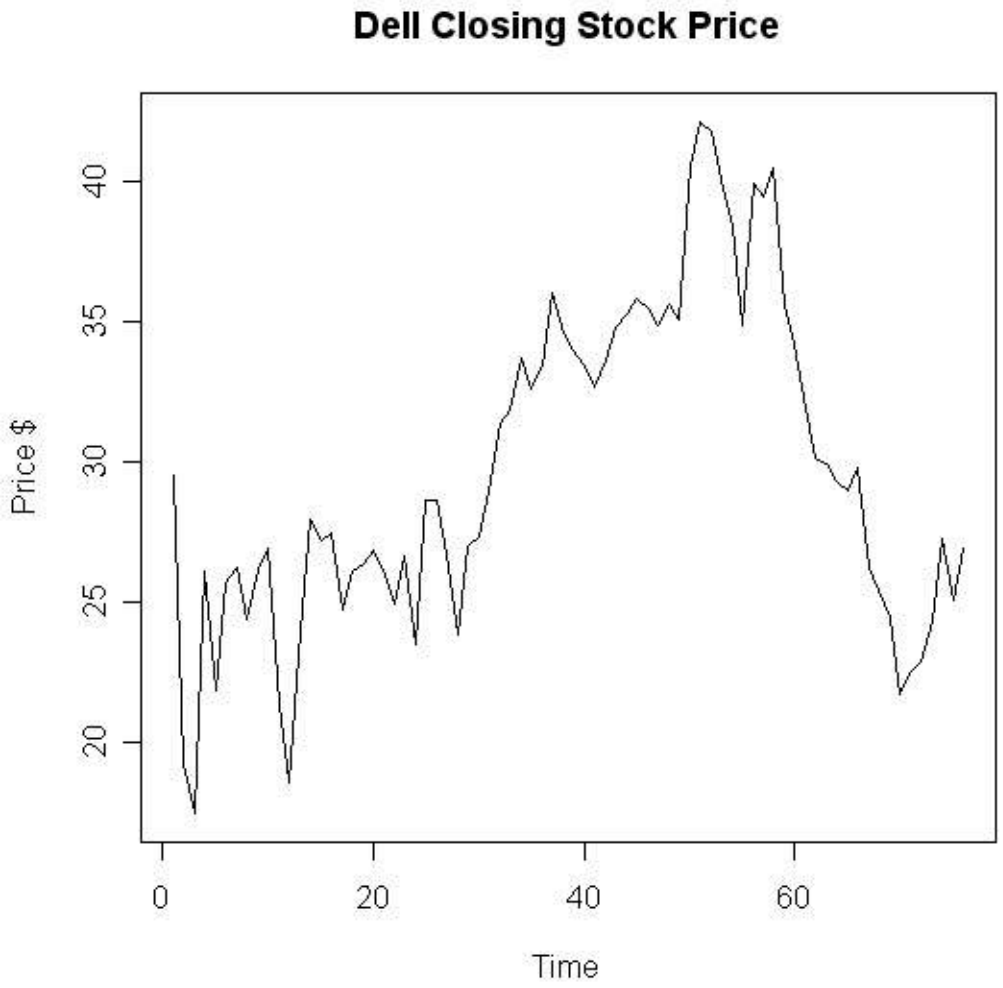

```
boxplot(dataset, col = rainbow(6), ylab = "Appropriate Units")
```

# Scatter plots -Plotting Two Vectors

```
plot(dataset$Hip, dataset$Waist,
  xlab = "Hip", ylab = "Waist",
main = "Circumference (in cm)", pch = 2, col = "blue")
```



# Line Plots




```
plot(t1,D2$DELL,type="l",main='Dell Closing Stock Price',
xlab='Time',ylab='Price $'))
```

# Adding a Legend

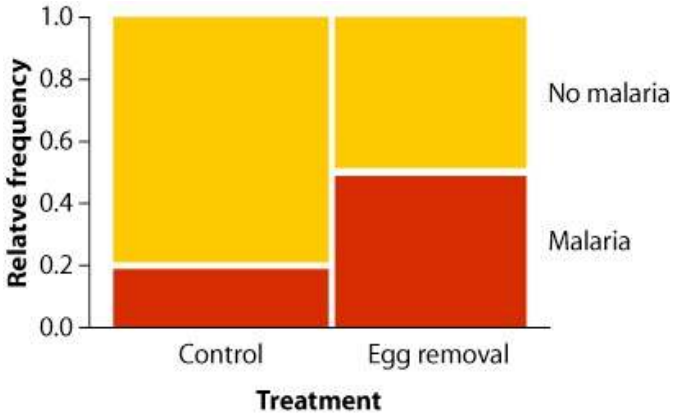

```
legend(60,45,c('Intel','Dell'),lty=c(1,2))
```

# Mosaic plot

## Association between reproductive effort and avian malaria

Table 2.3A. Contingency table showing incidence of malaria in female great tits subjected to experimental egg removal.

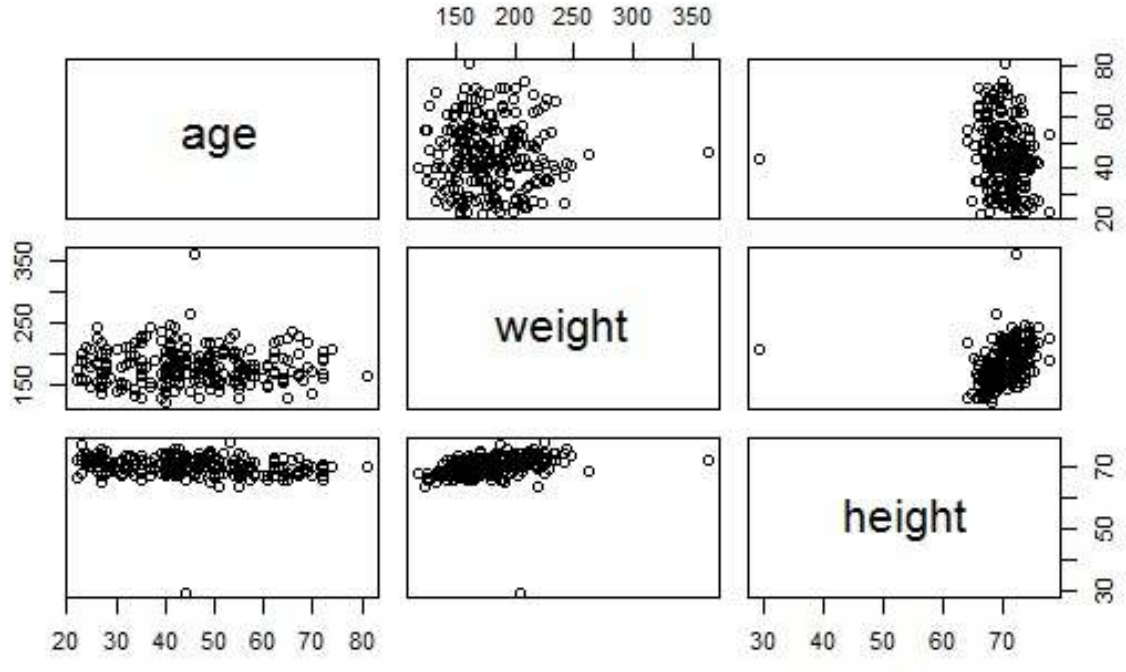
| | control group | egg removal group | row total |
|---|---|---|---|
| malaria | 7 | 15 | 22 |
| no malaria | 28 | 15 | 43 |
| column total | 35 | 30 | 65 |




```
>library(vcd)
>mosaic(HairEyeColor, shade=TRUE, legend=FALSE)
```

# Plotting Contents of a Dataset as Matrix

>plot(dataset[c(5,6,7)])

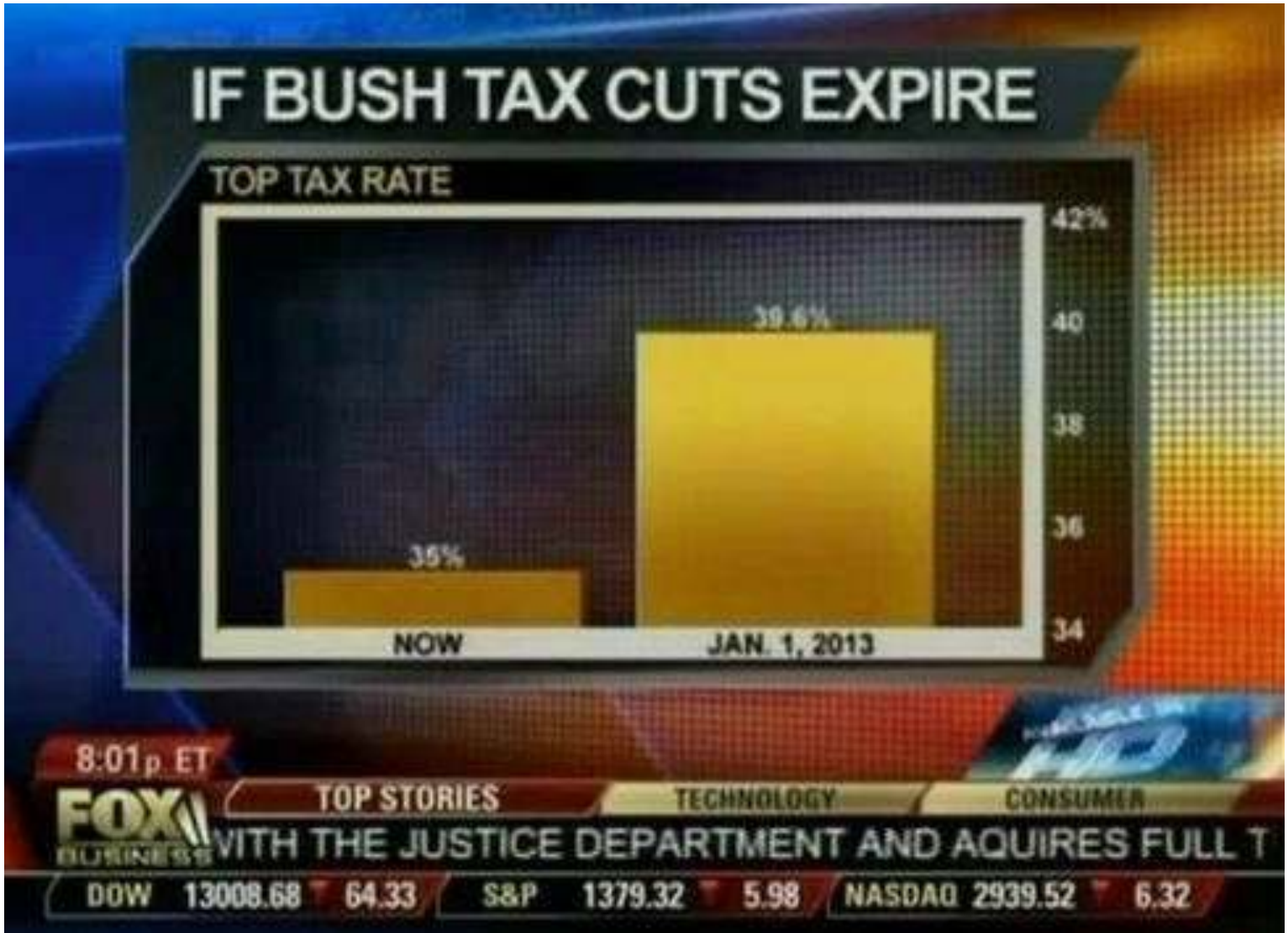


# Effective Visualizations

1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color strategically
5. Tell a story with data

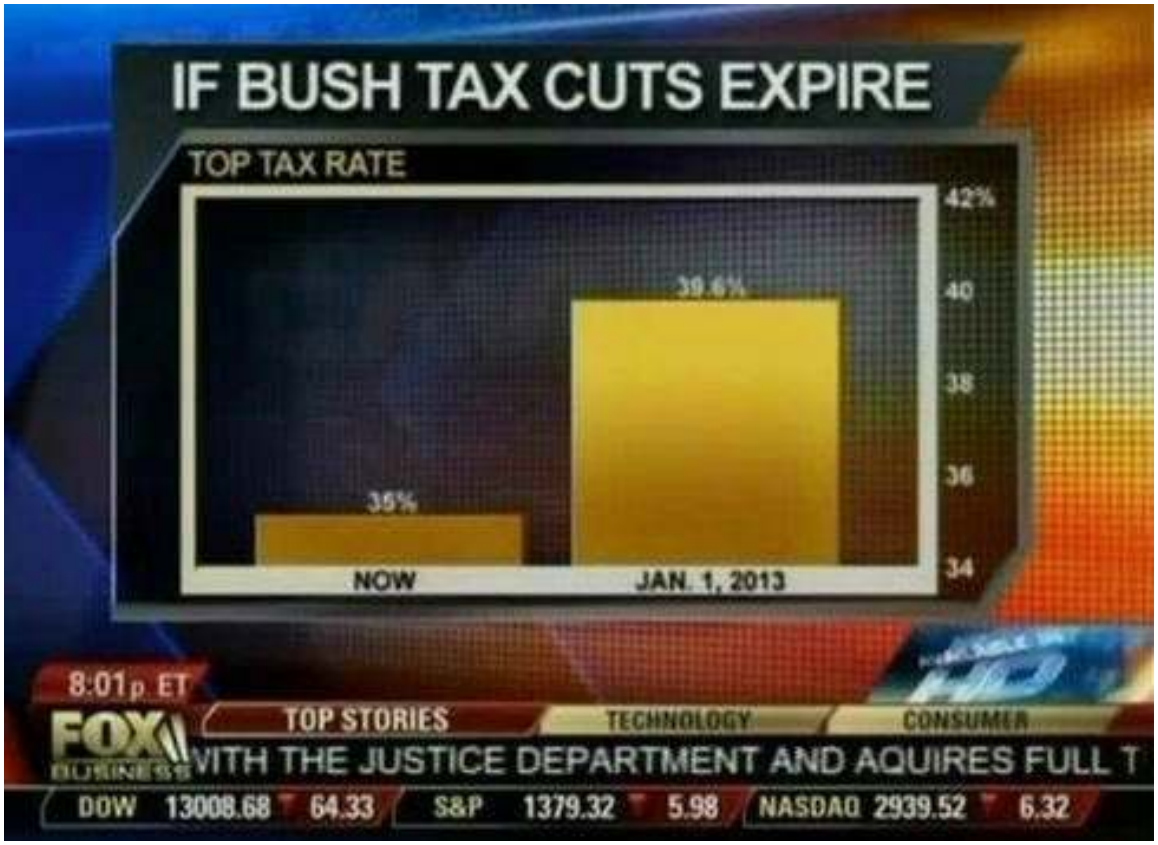
# Graphical Integrity

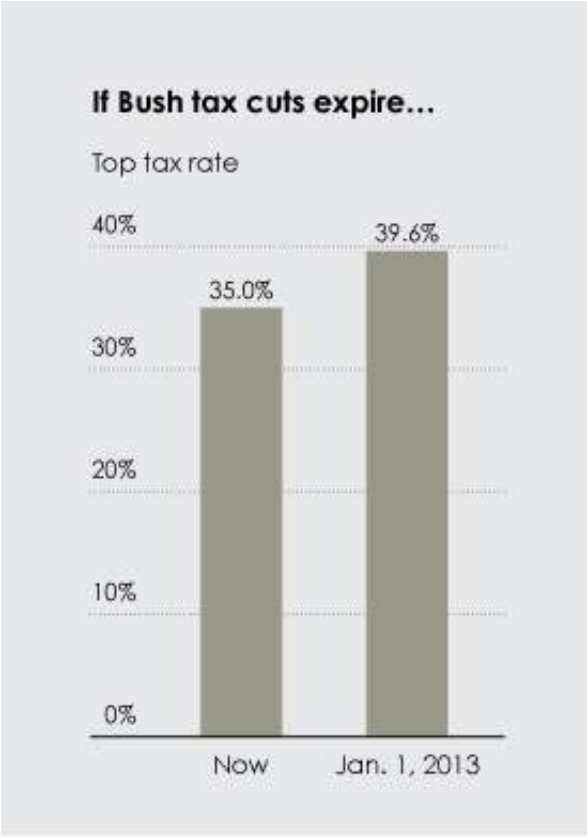
# Graphical Integrity




Flowing Data

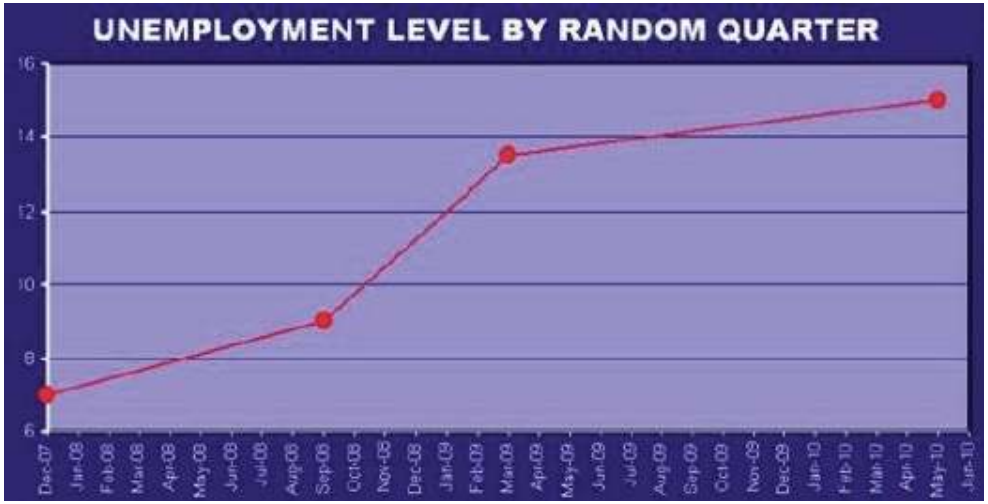
# Scale Distortions



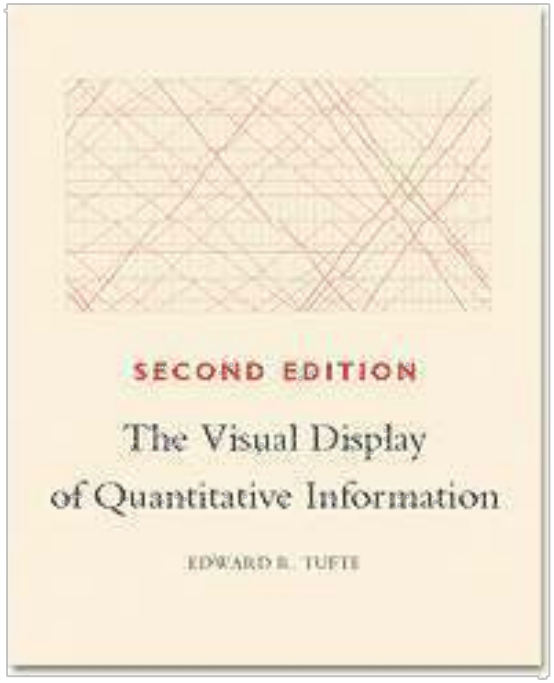

JOB LOSS BY QUARTER
15 MIL
13.5 MIL
9 MIL
7 MIL
DEC '07
SEPT '08
MARCH '09
JUNE '10
FOX NEWS .COM
SOURCE: BLS
AMERICA'S NEWSROOM

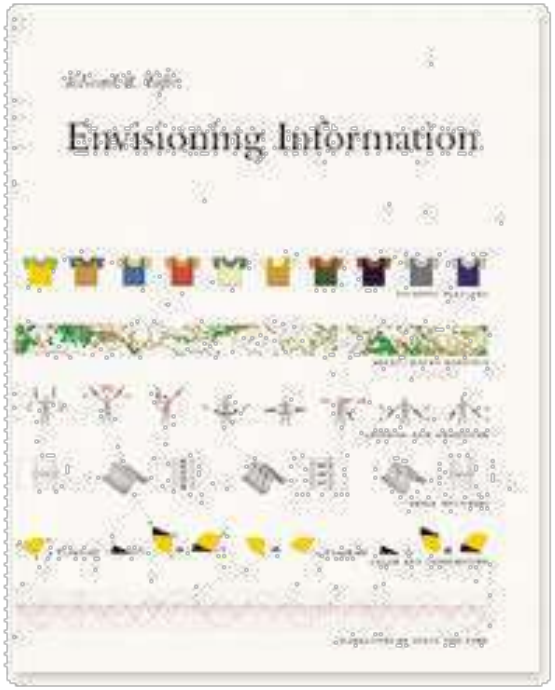
# Scale Distortions

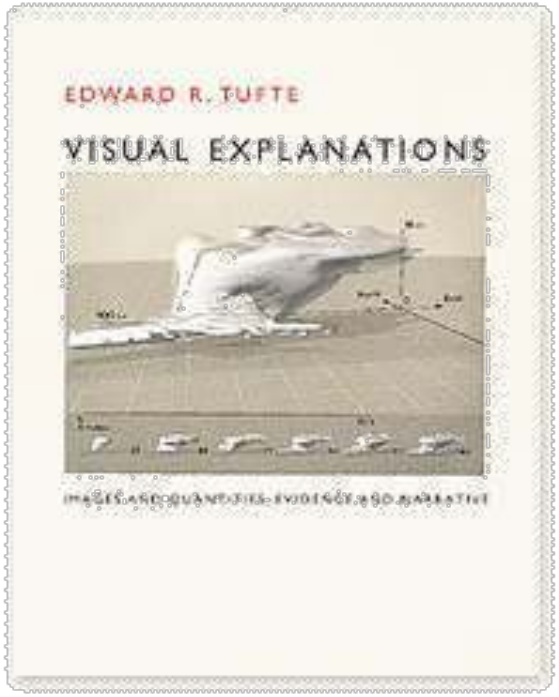



# Keep It Simple

# Edward Tufte

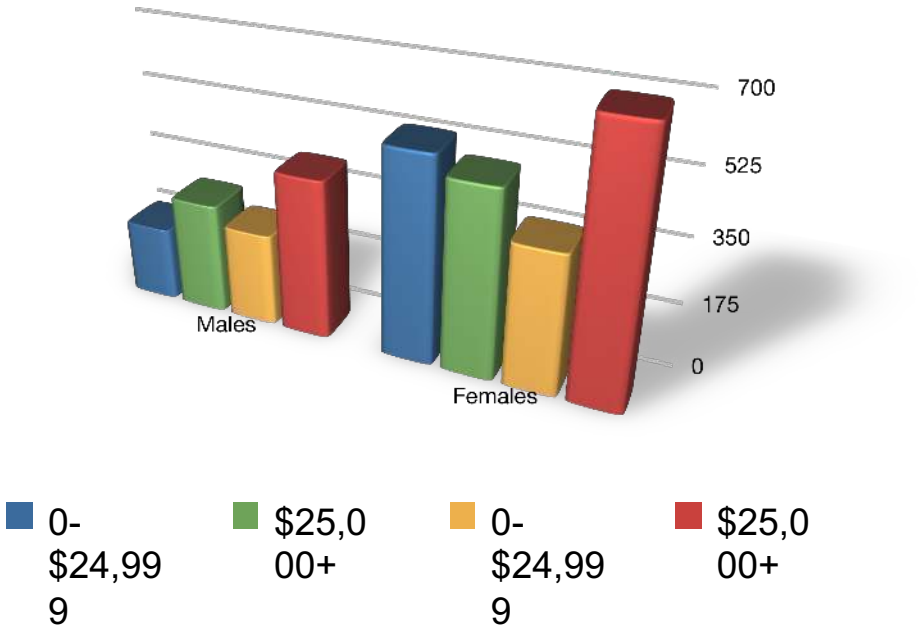


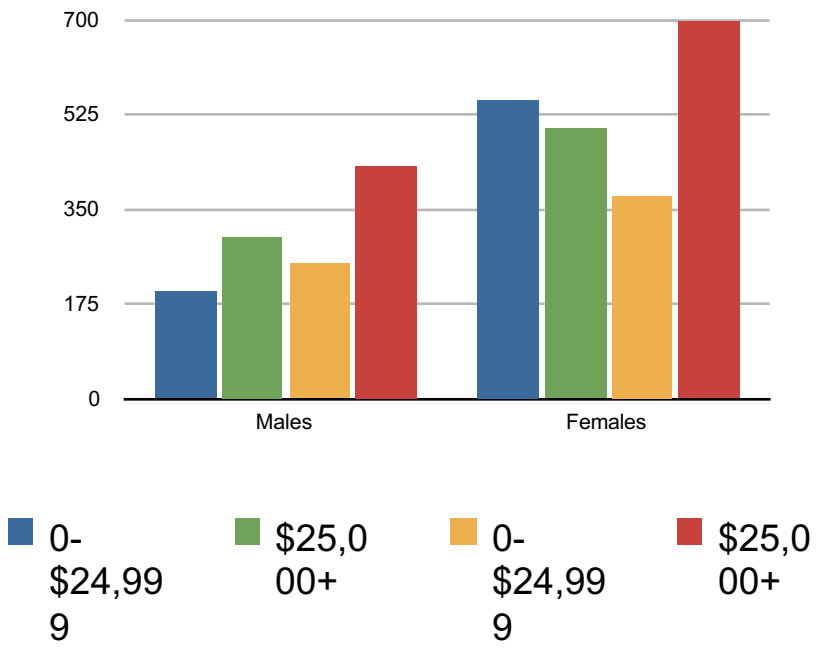
# Maximize Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

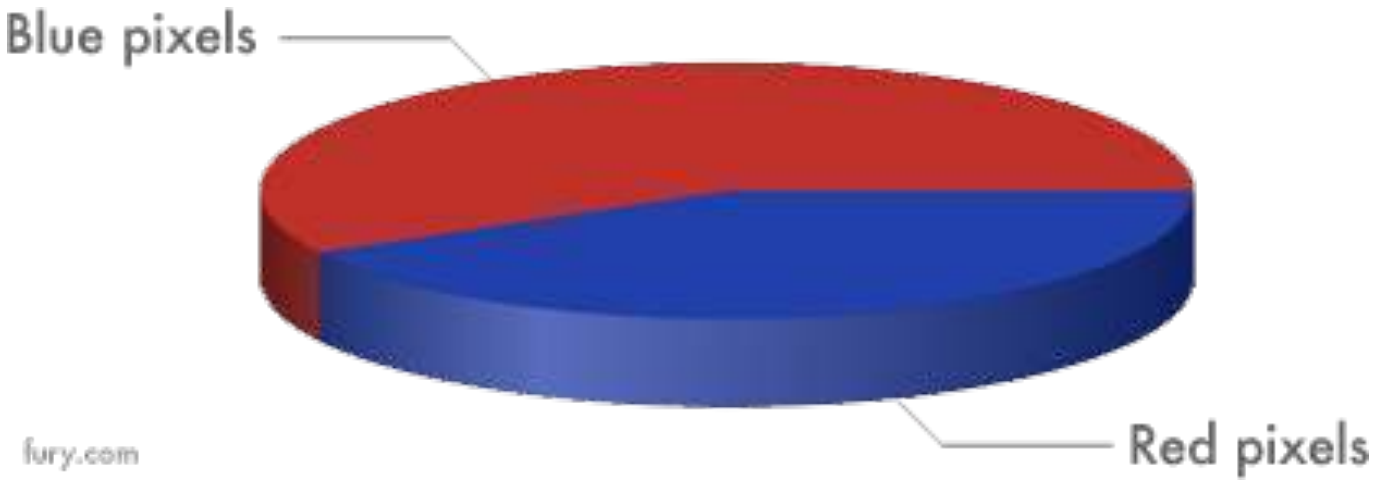


# Maximize Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$



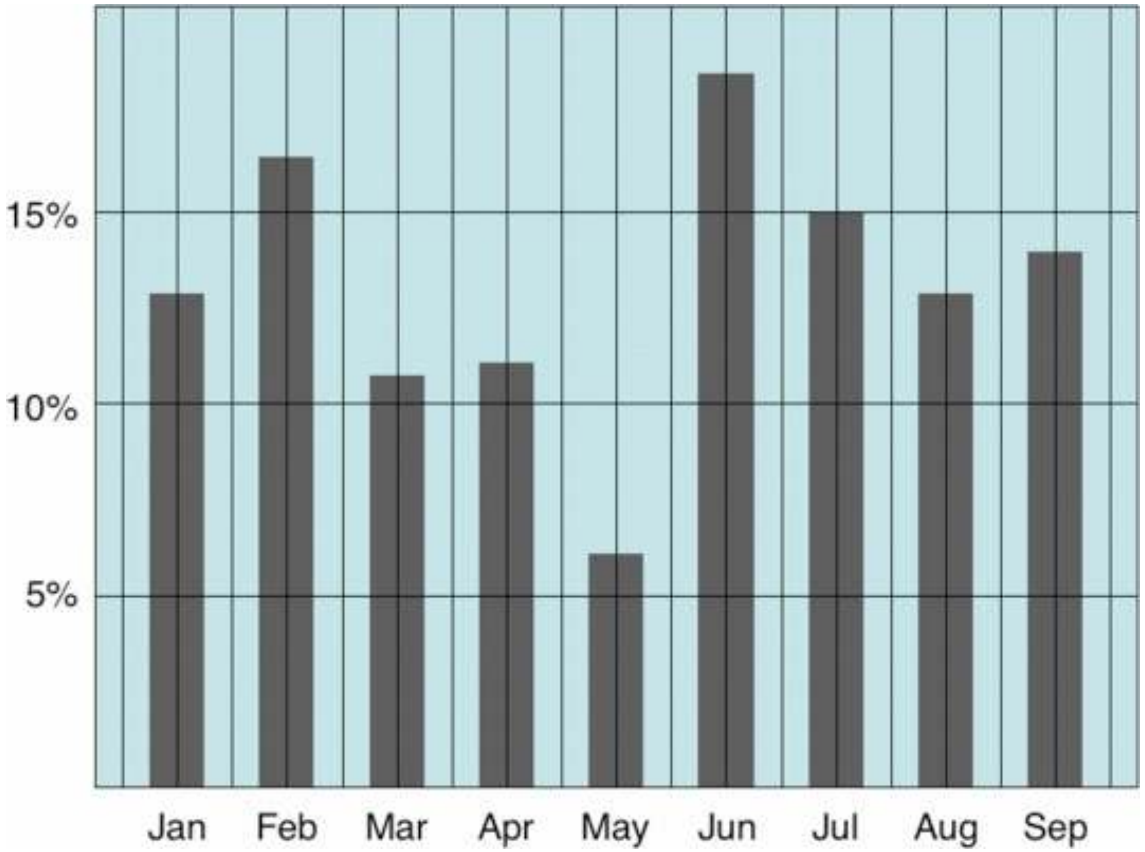
# Why 3D pie charts are bad



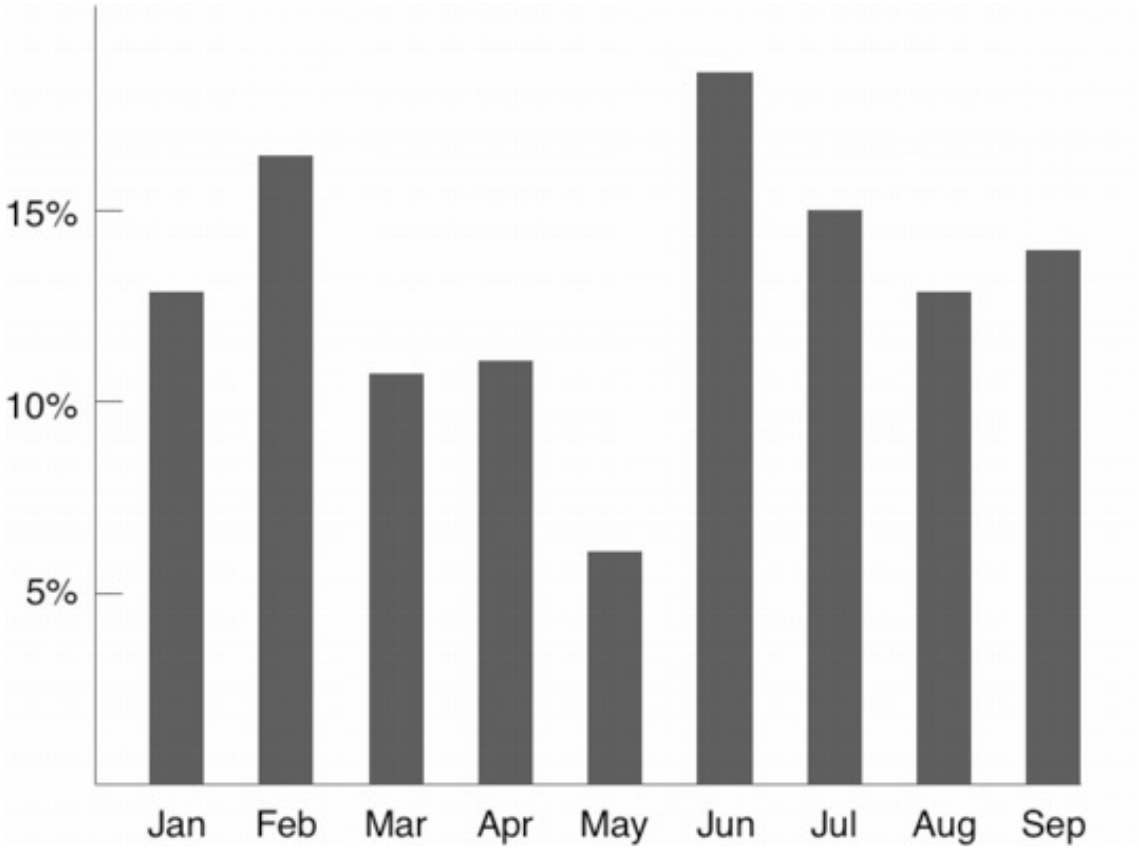

Kevin Fox

# Avoid Chartjunk

Extraneous visual elements that distract from the message

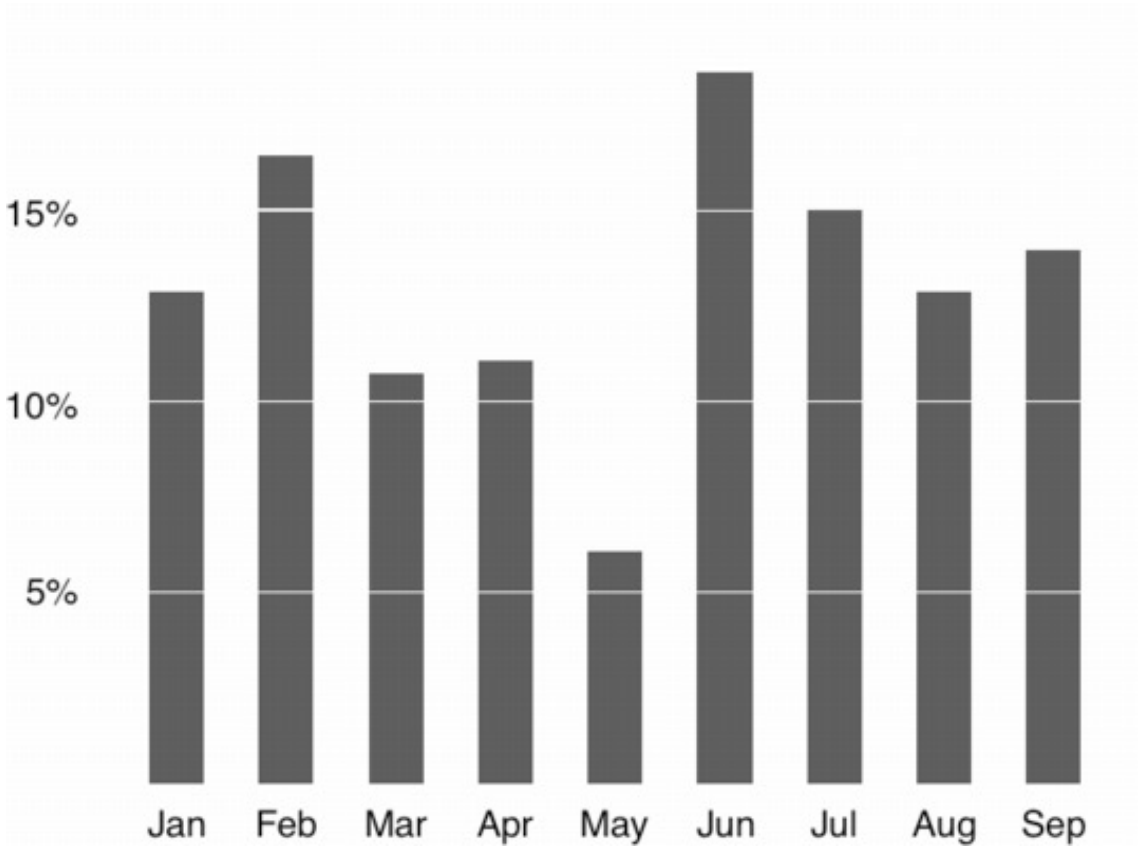


# Avoid Chartjunk



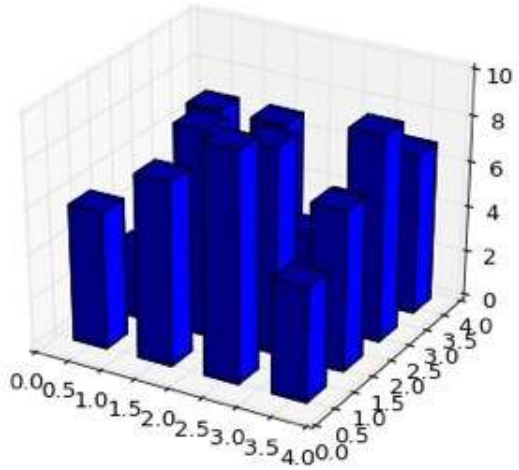

ongoing,Tim Brey

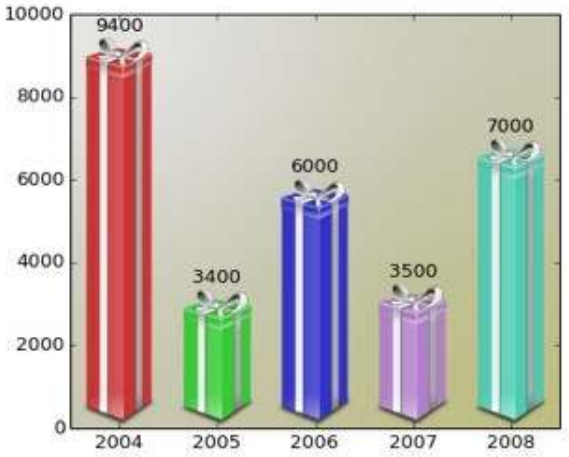
# Avoid Chartjunk



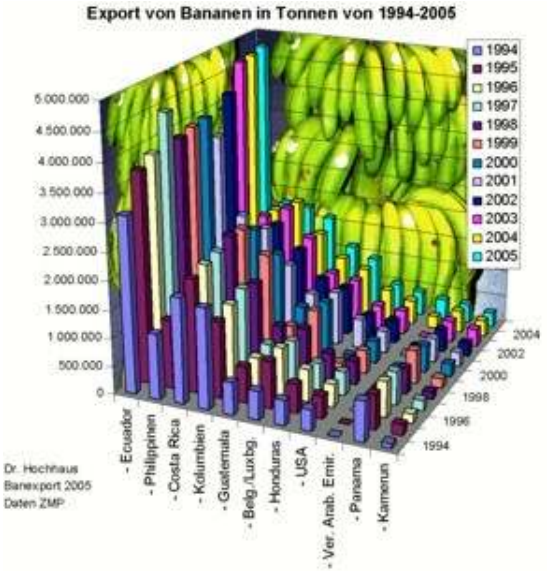

ongoing,Tim Brey

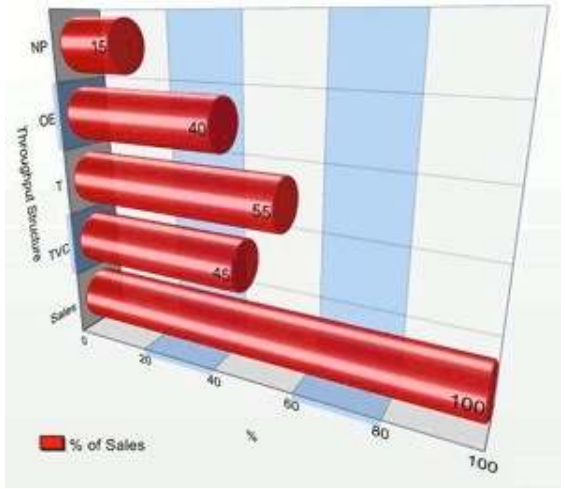
# Don’t!




matplotlib gallery

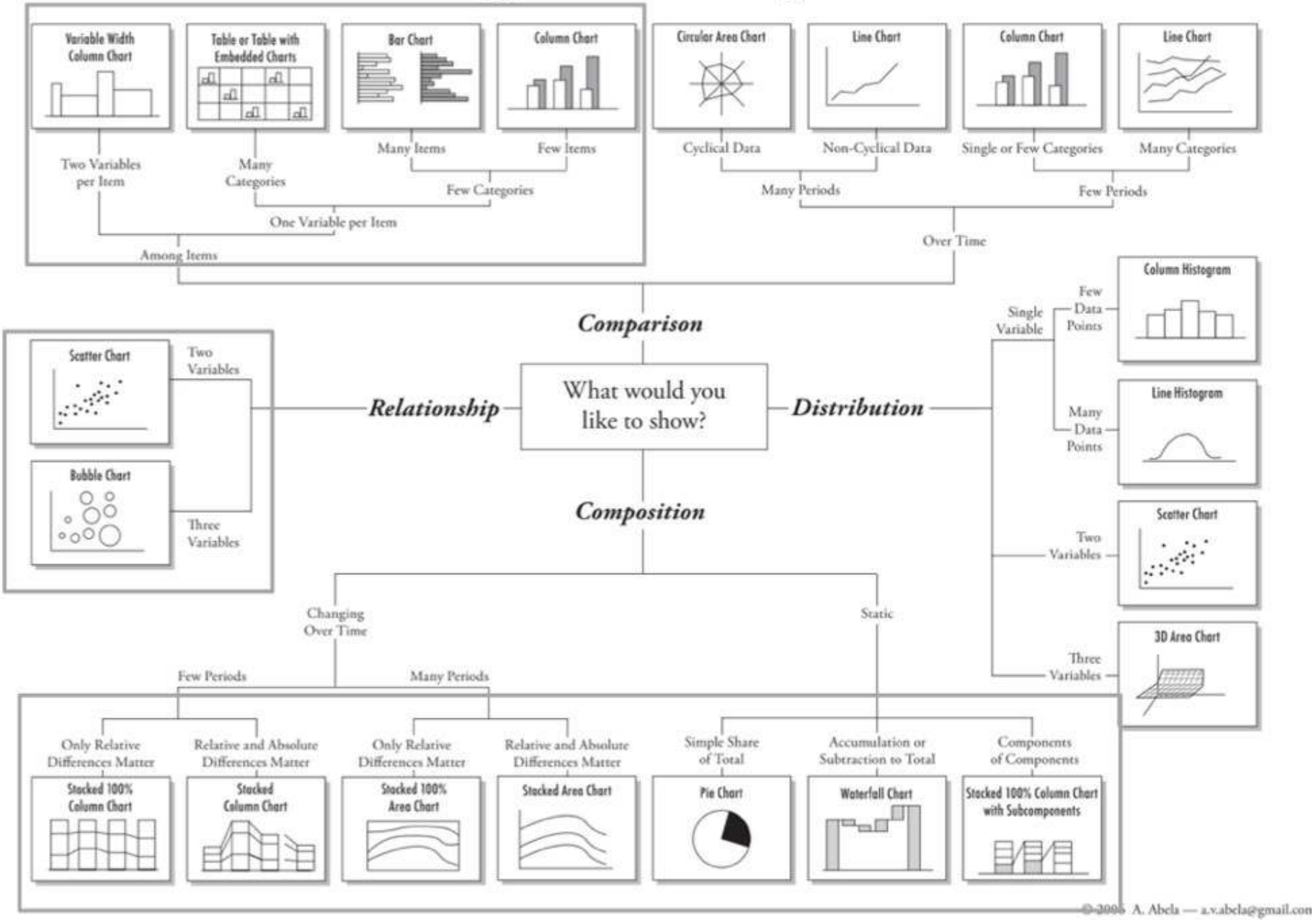



Excel Charts Blog

# Use The Right Display

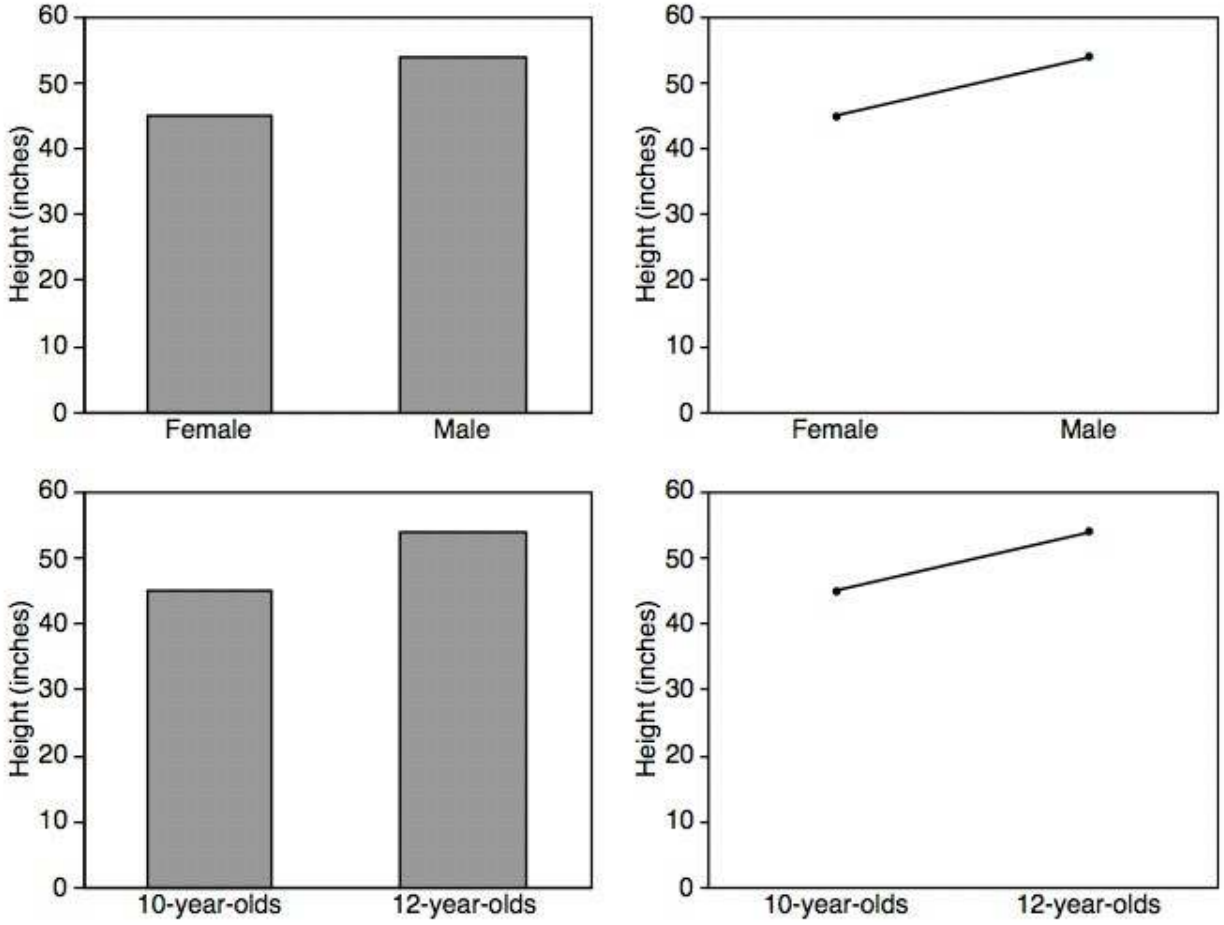
# Chart Suggestions—A Thought-Starter



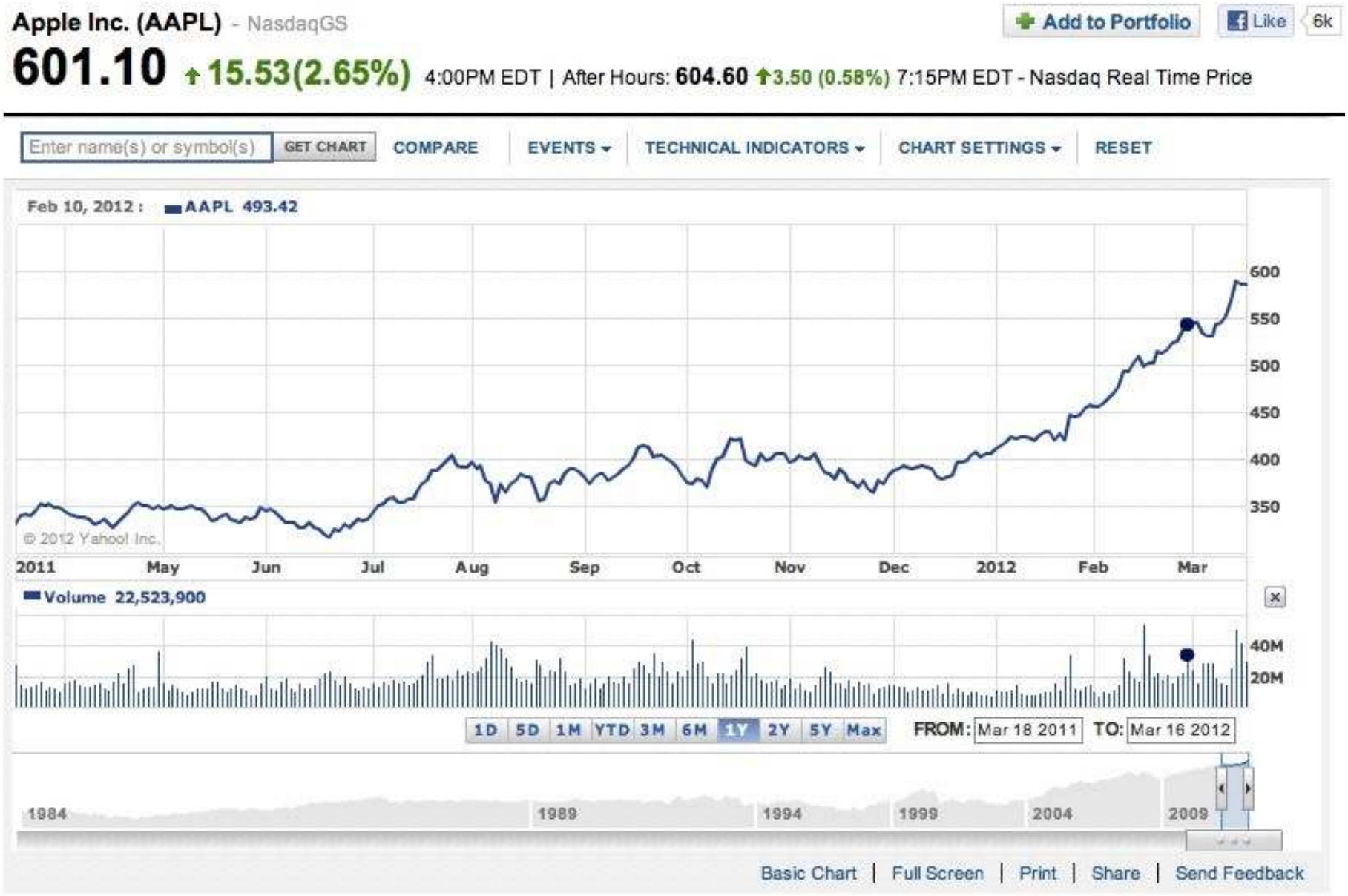

http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

# Comparisons

# Bars vs. Lines



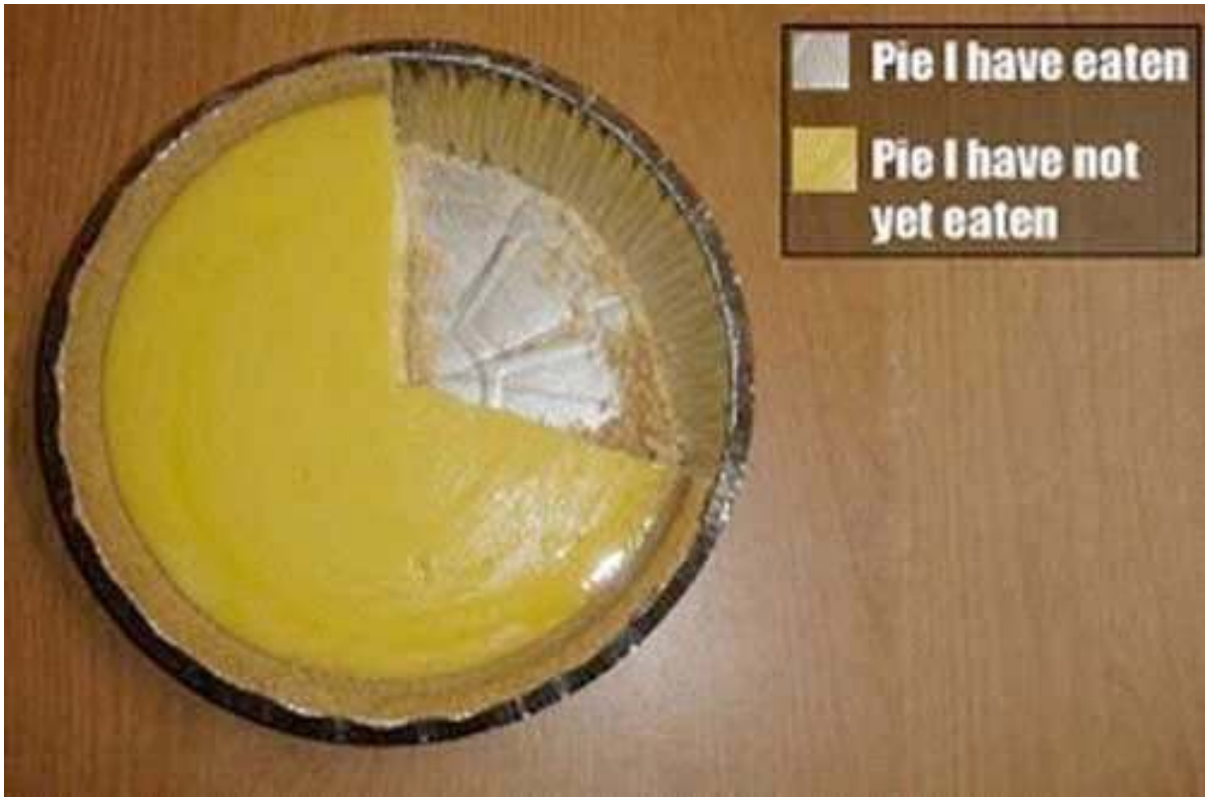
# Trends



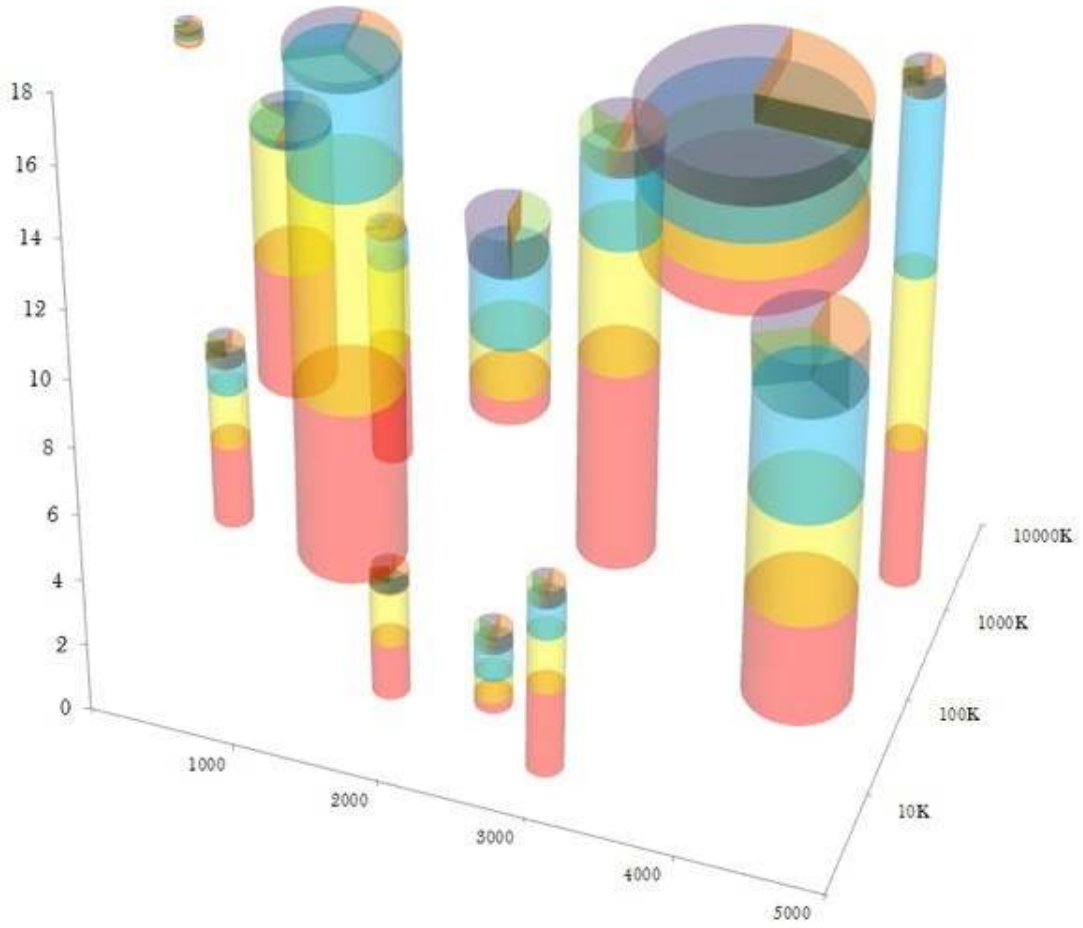

Yahoo! Finance

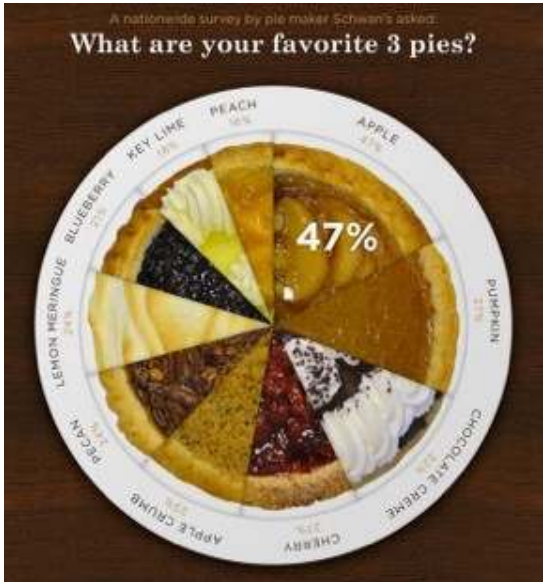
# Proportions

# Pie Charts

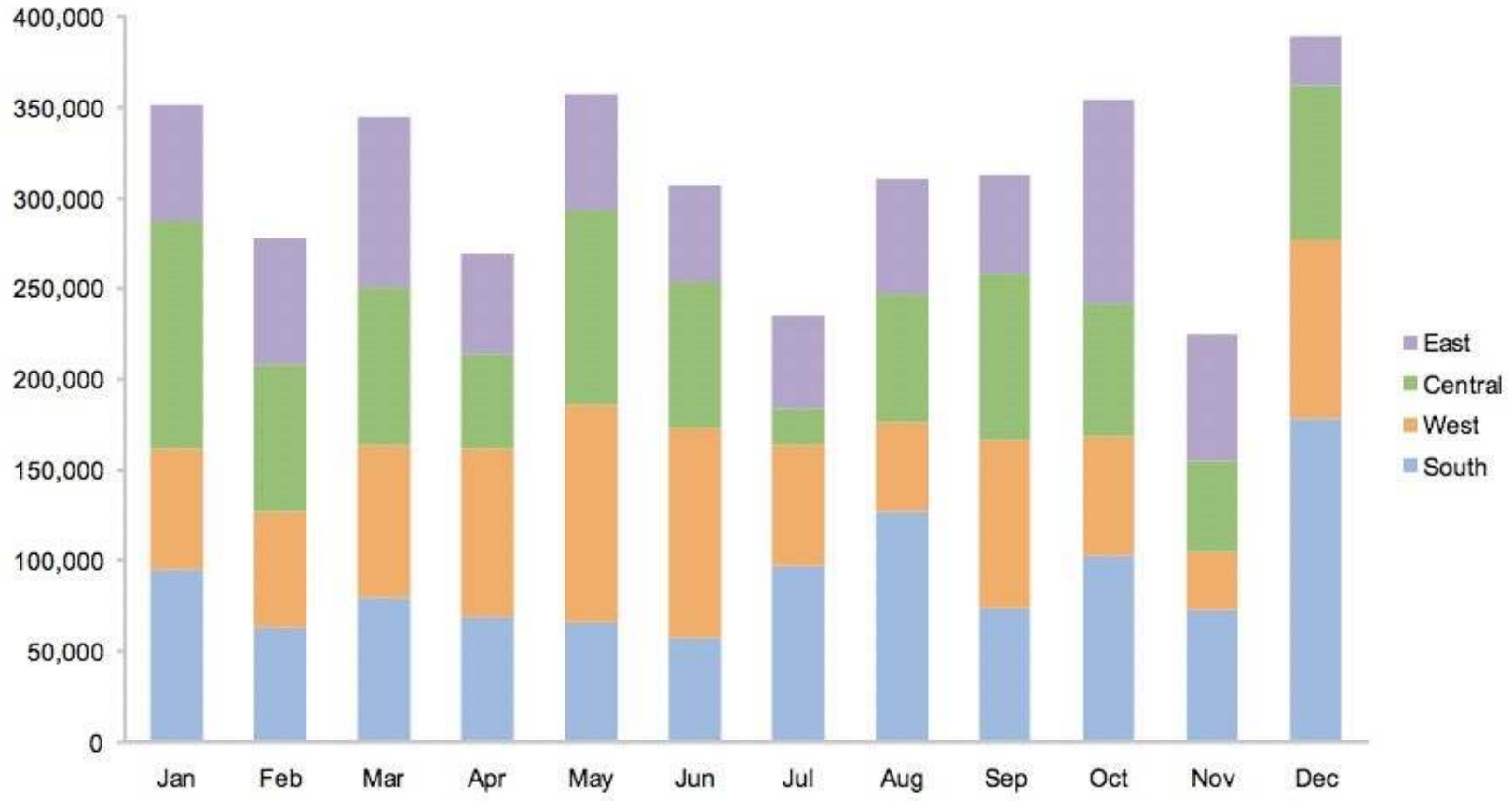


# eagerpies.com



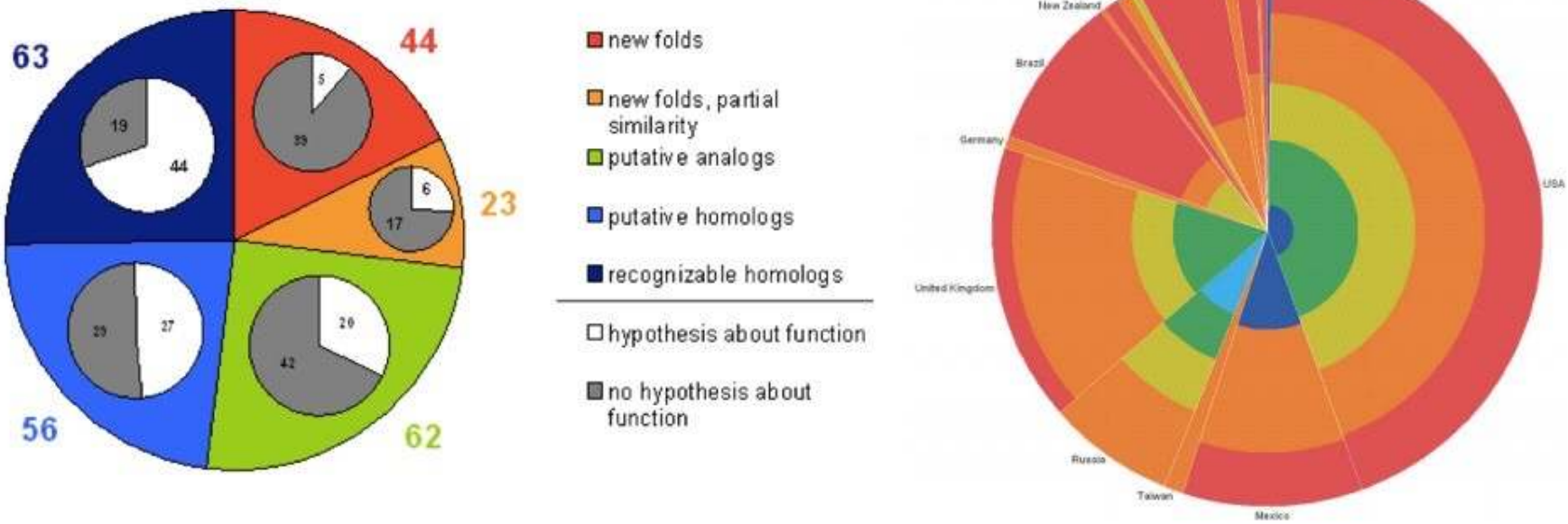




# Stacked Bar Chart



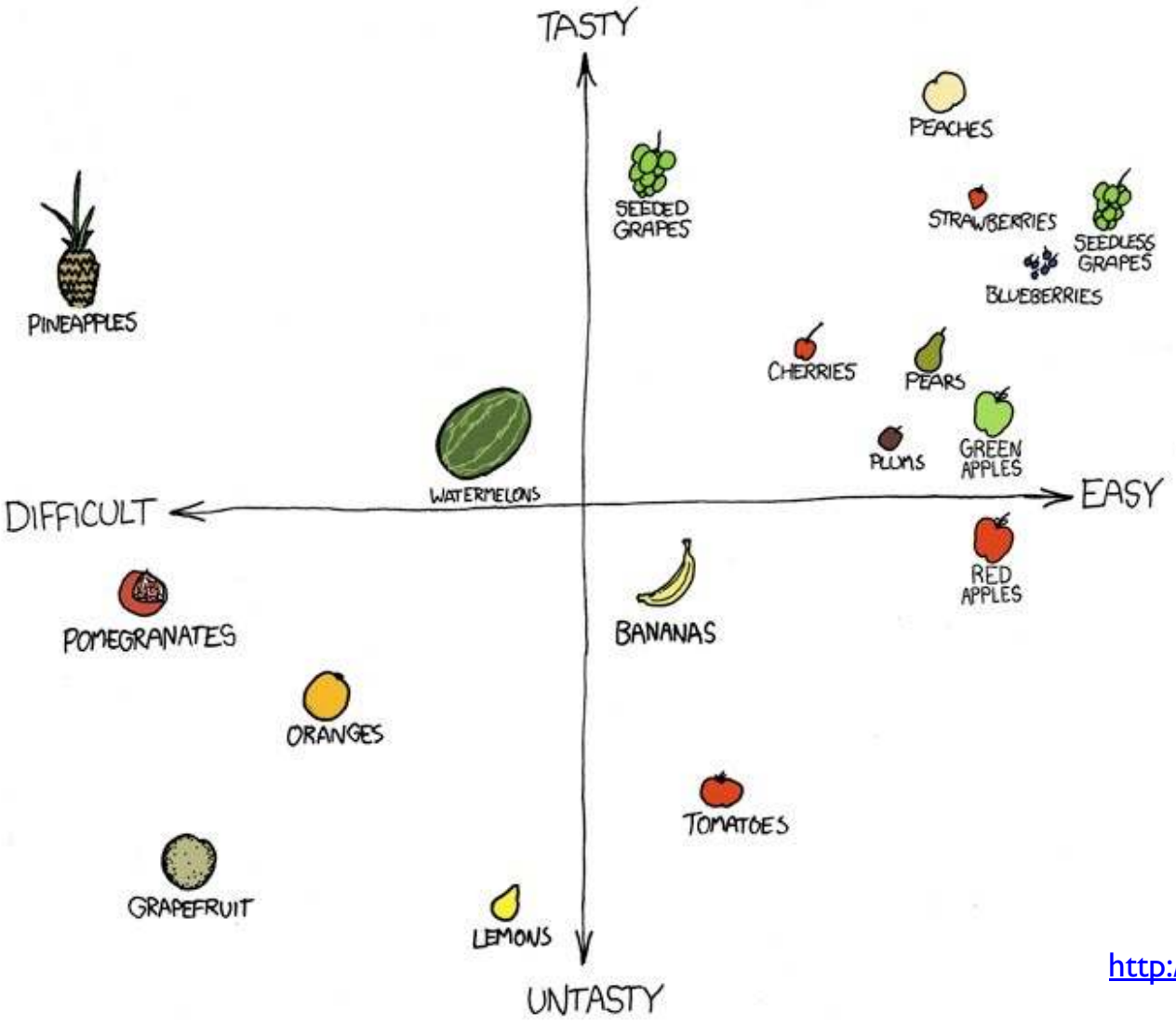

S.Few

# Don't!


63
44
23
56
62
19
44
5
39
6
17
29
27
20
42
new folds
new folds, partial similarity
putative analogs
putative homologs
recognizable homologs
hypothesis about function
no hypothesis about function
Amount per Country
USA
US
Italy
Norway
New Zealand
Brazil
Germany
United Kingdom
Russia
Taiwan
Mexico

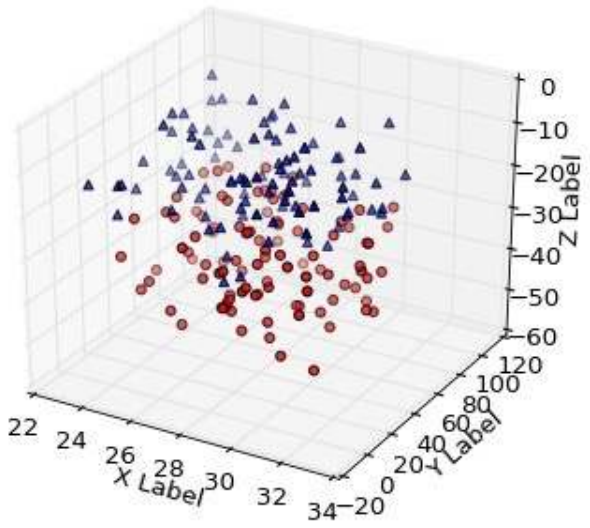
# Correlations

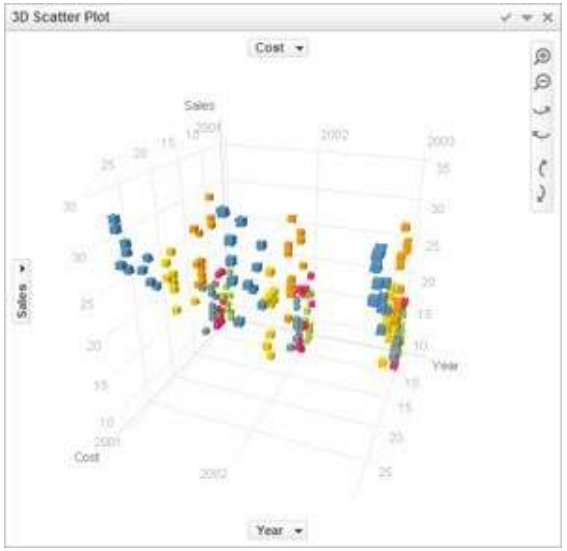
# Scatterplots

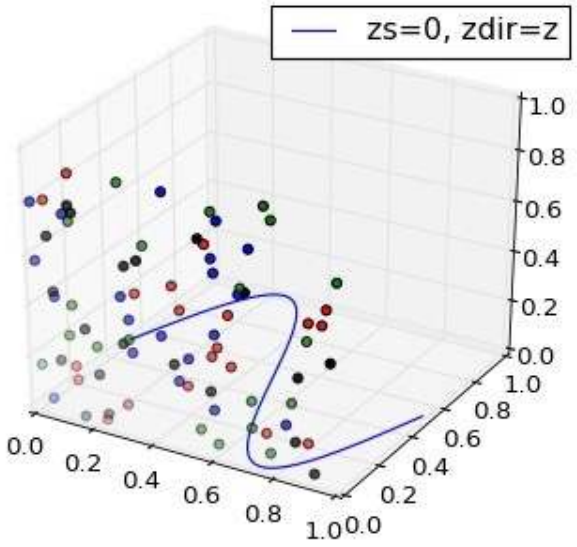



http://xkcd.com/388/

# Don't!

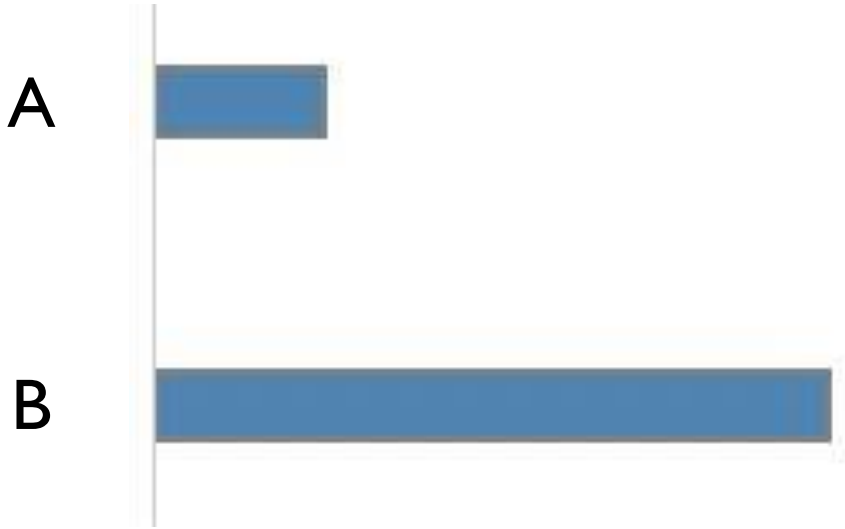

X Label
Y Label
Z Label


3D Scatter Plot
Cost
Sales
Year

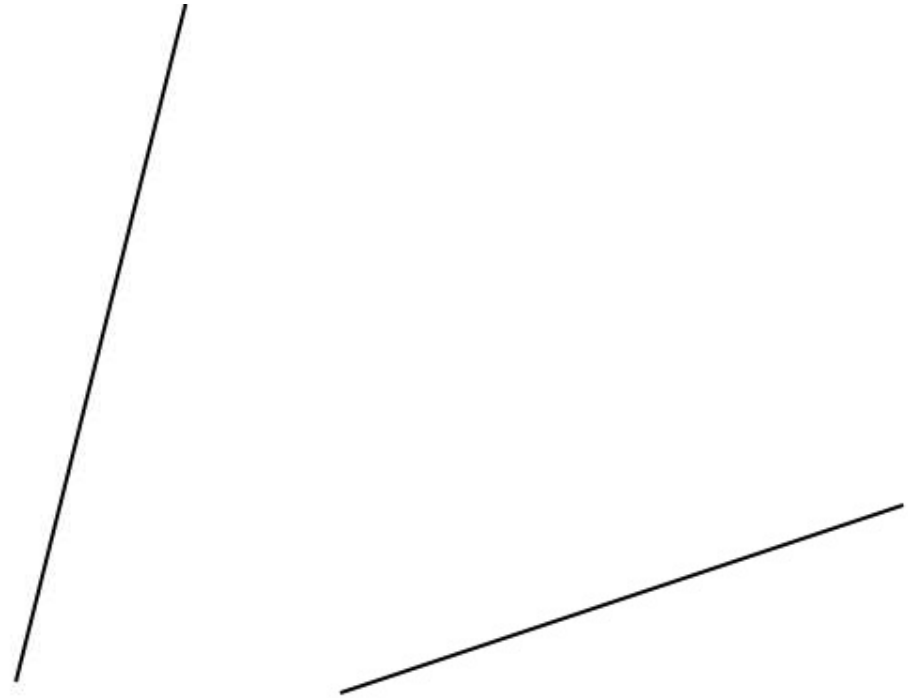

zs=0, zdir=z

# Perceptual Effectiveness

# How much longer?



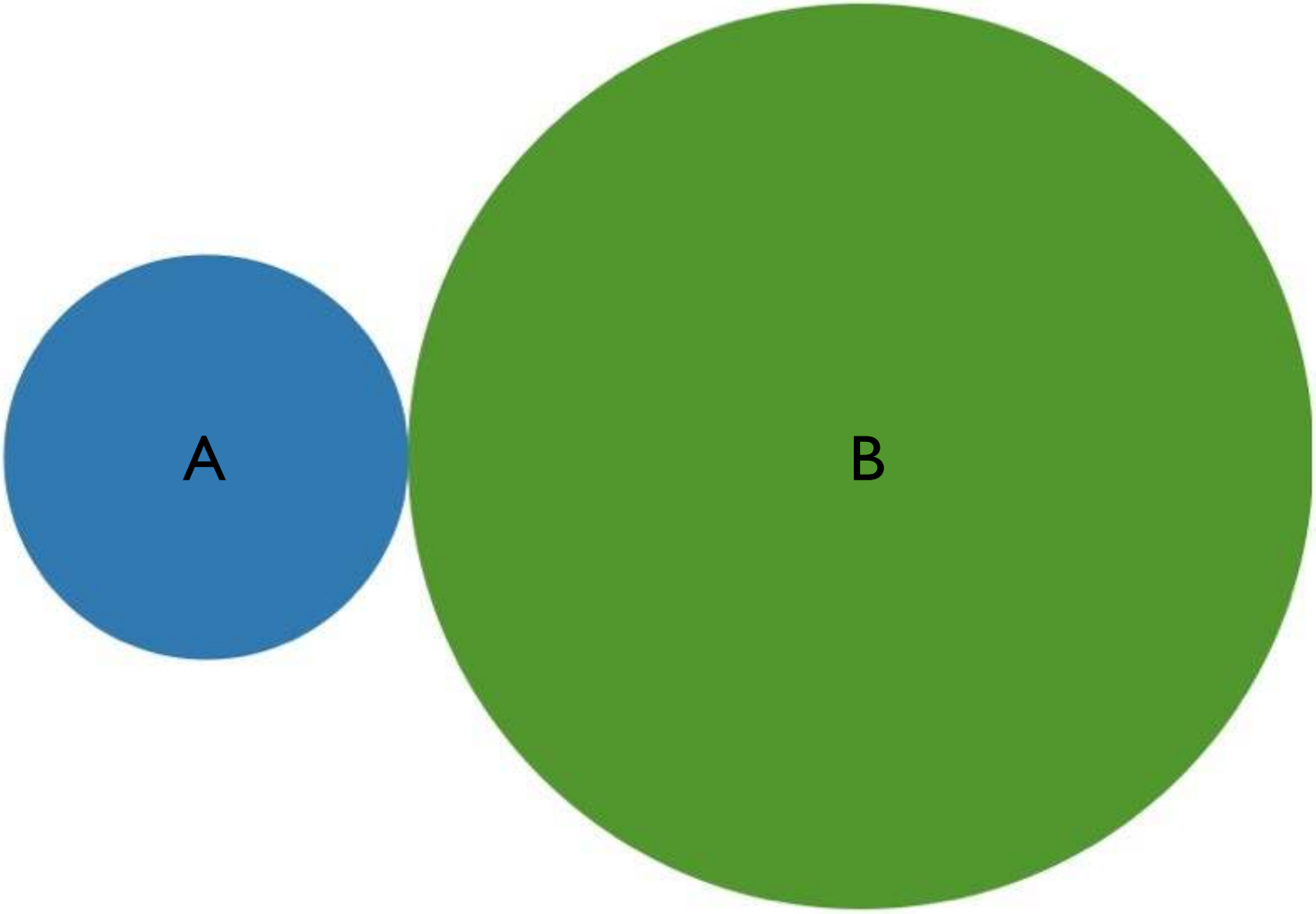
# How much steeper slope?



A

B

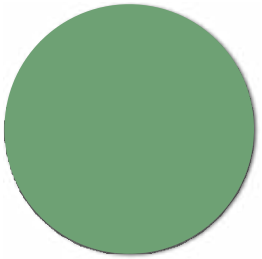
# How much larger area?



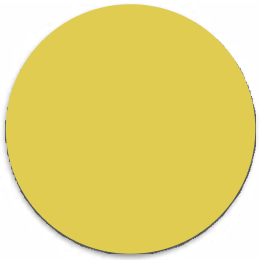
# How much darker?

A B

# How much bigger value?

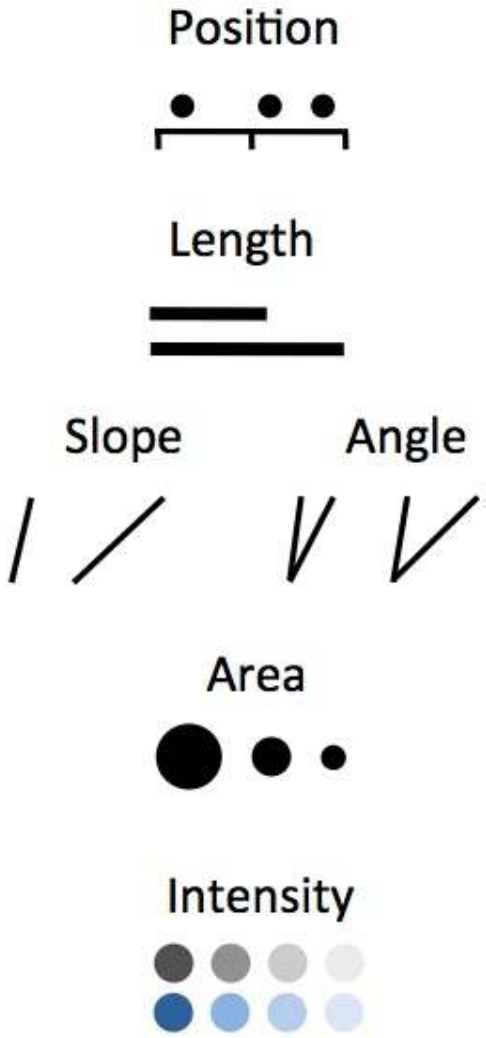


A



B

Most Efficient




Least Efficient

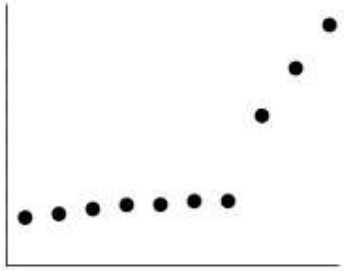
Quantitative

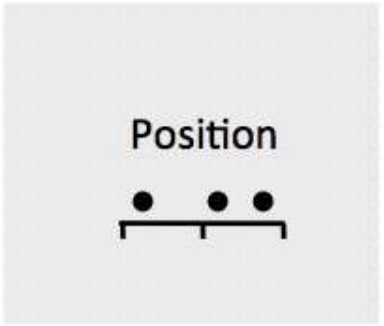
Ordered

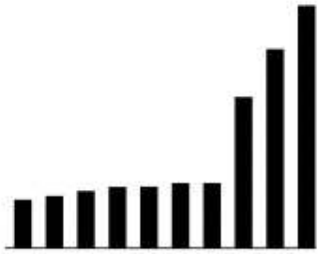
Categories

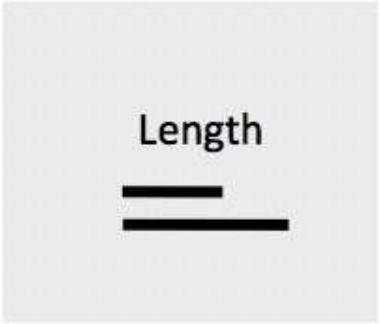
C. Mulbrandon
VisualizingEconomics.com

# Most Effective

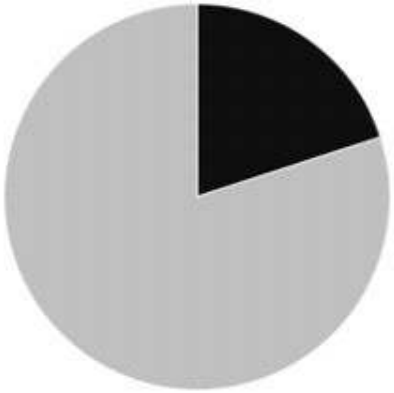


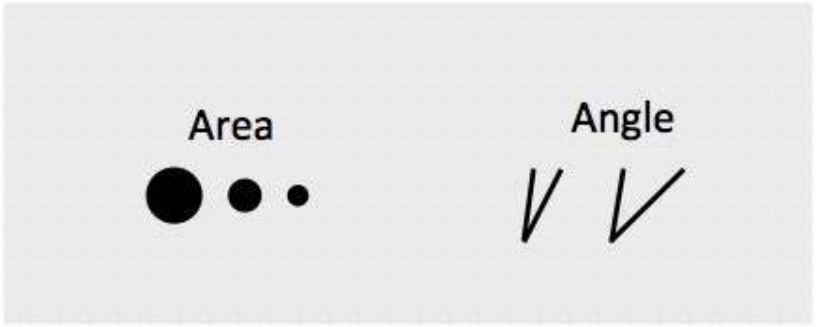


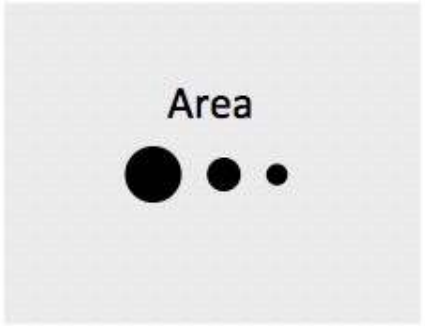





# Less Effective

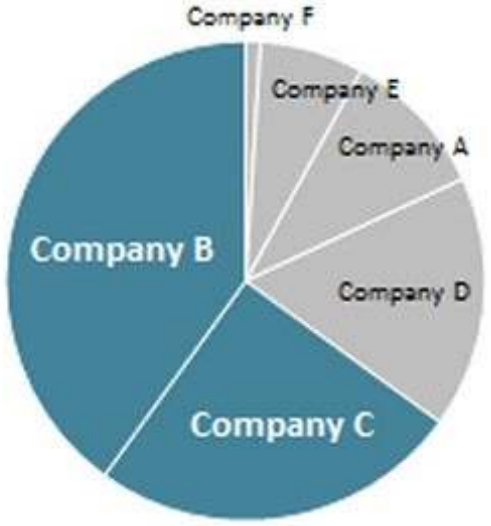


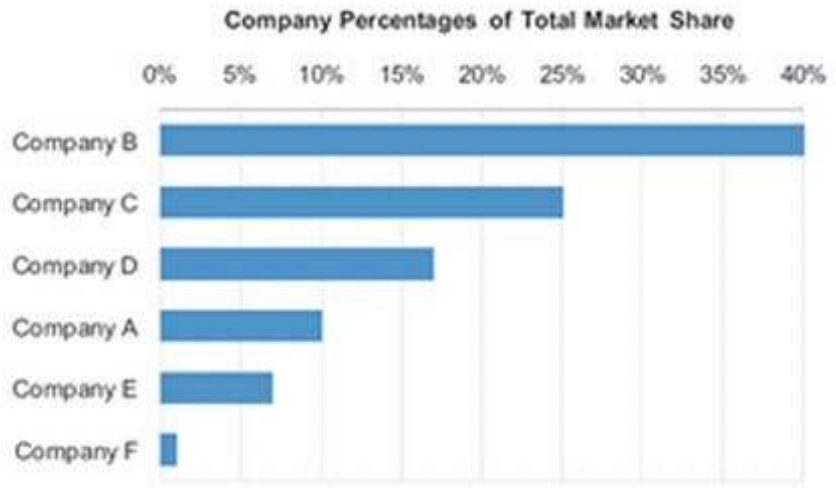


# Pie vs. Bar Charts

65% of the market is controlled by companies B and C



# Least Effective

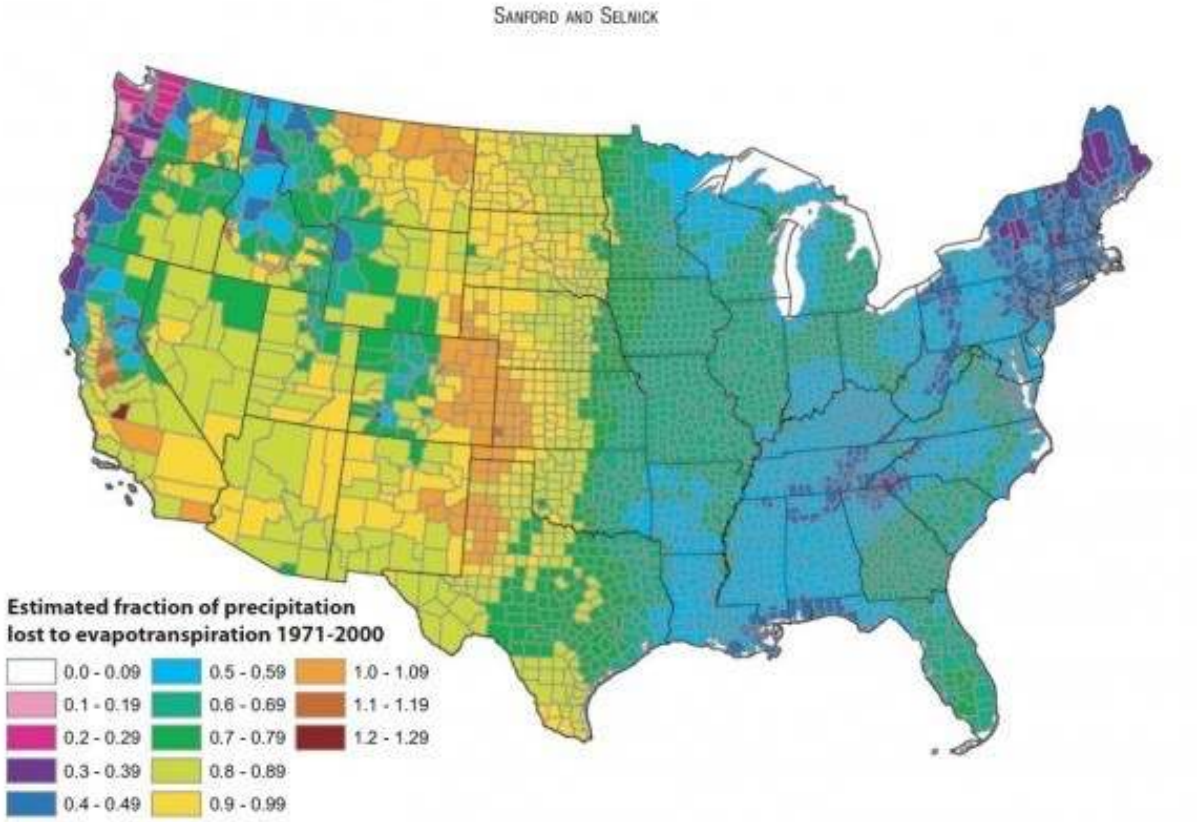
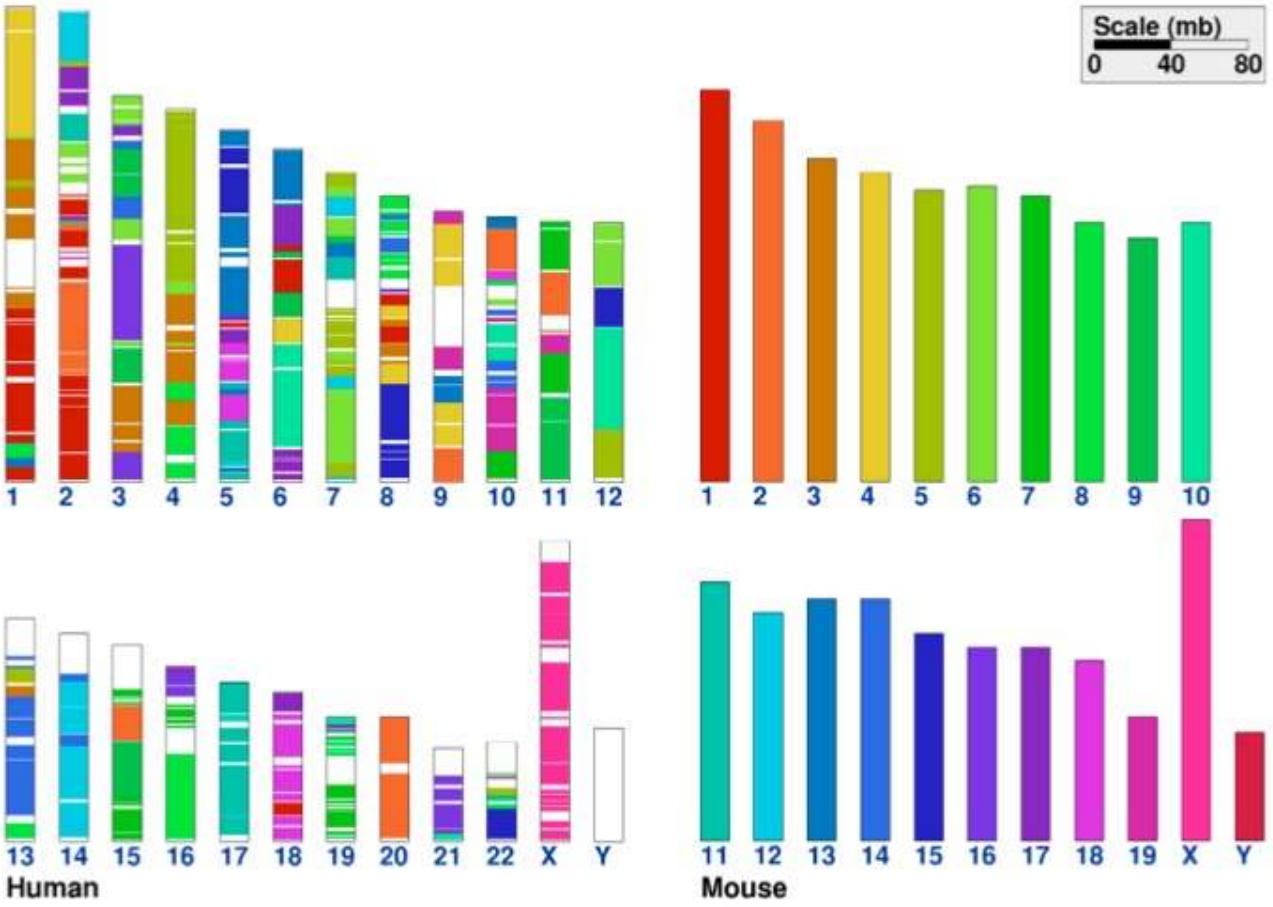



FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation ($P$) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of $ET/P$ were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

Cliff Mass

# Use Color Strategically

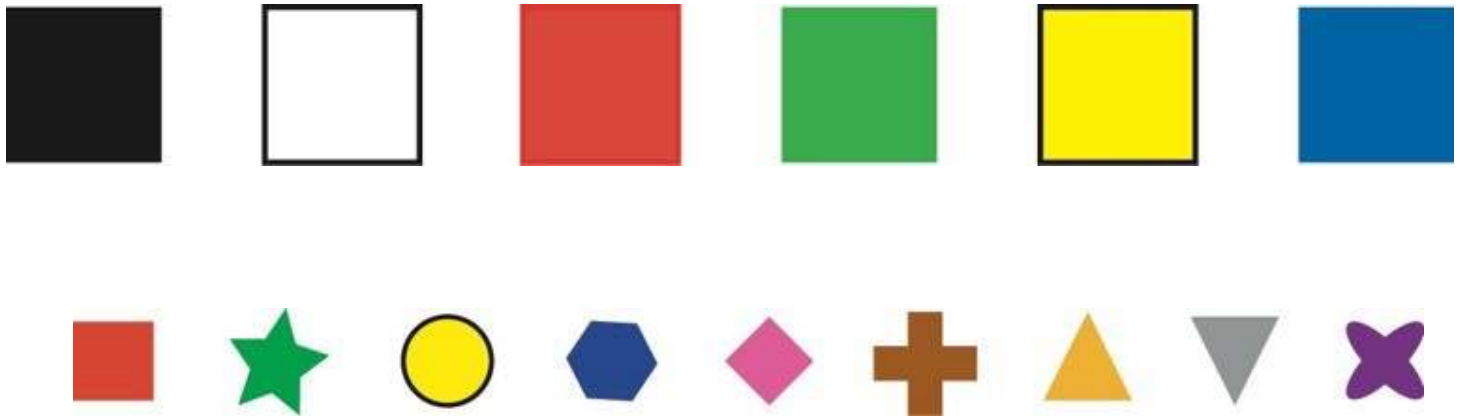
# Color Discriminability




Sinha 2007

# Colors for Categories

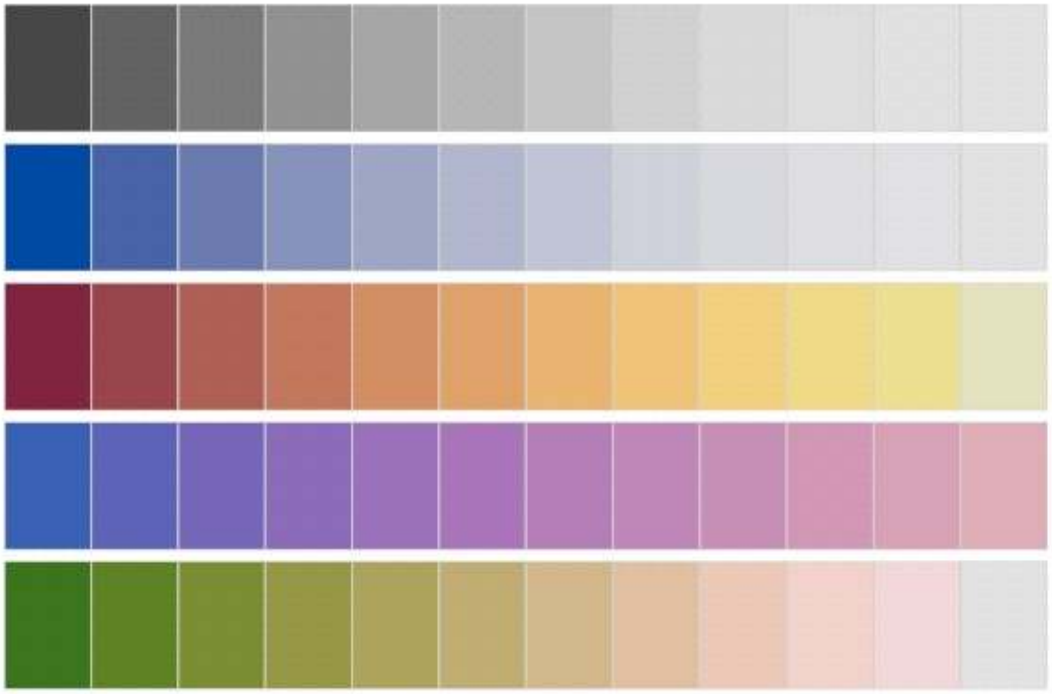
Do not use more than 5-8 colors at once



Ware, "Information Visualization"

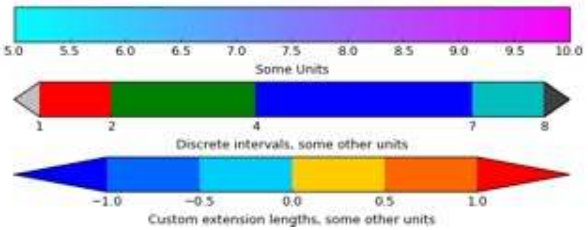
# Colors for Ordinal Data

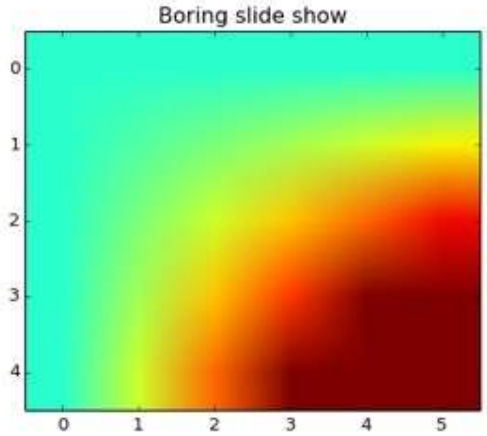
## Vary luminance and saturation

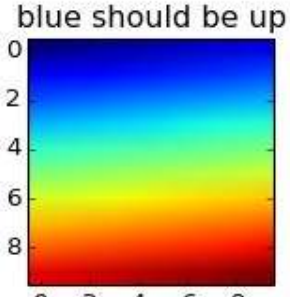


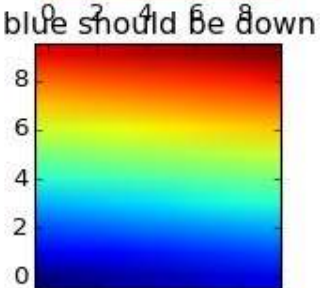
Zeilis et al, 2009, "Escaping RGBland: Selecting Colors for Statistical Graphics"

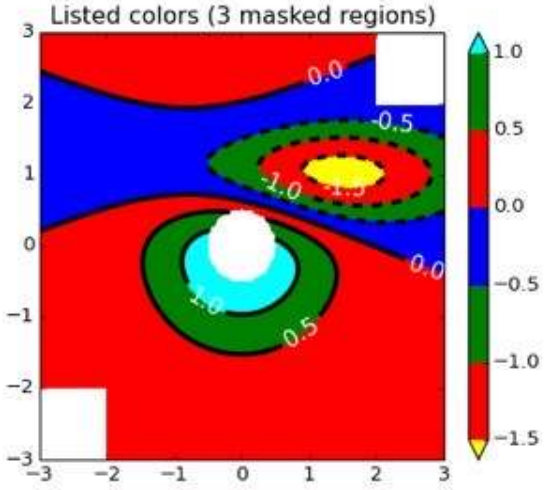
# Avoid Rainbow Colors!

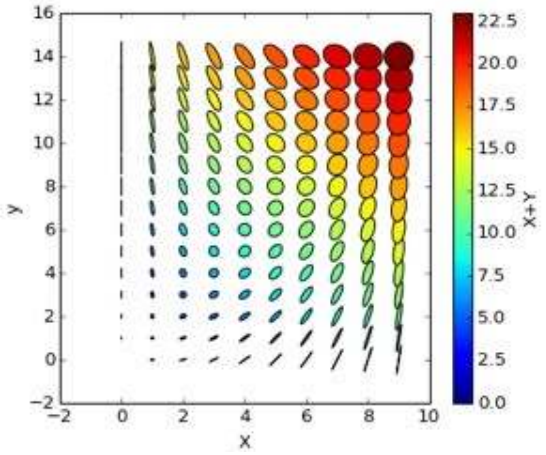


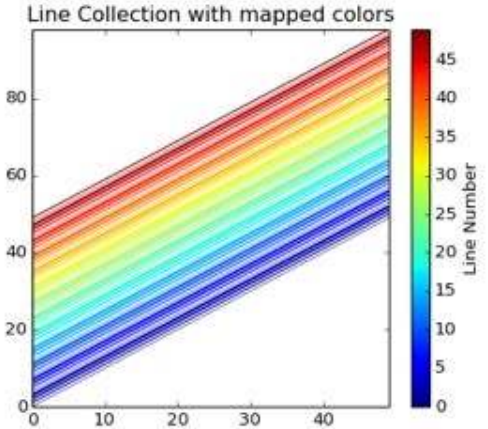




Perceptually nonlinear

matplotlib gallery

# Color Blindness

Protanope

Deuteranope

Tritanope

Red / green deficiencies

Blue /Yellow deficiency

Based on slide from Stone