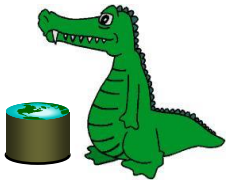
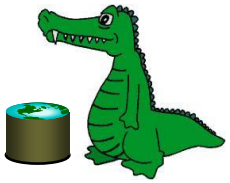


Logistics

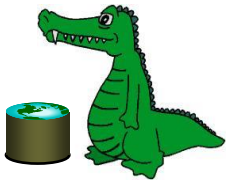
- Lab 2 material is posted
  - Basic Statistics + Visualizations
  - In-class lab and quiz on Monday
  - Homework due Monday midnight

- Lab 1 scores will be posted today/tomorrow
  - Regrade policies in syllabus - Need to submit written explanation



Review

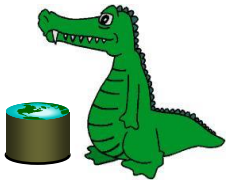
- Random Variables and Sample Statistics
  - Population vs. Sample
  - Probability, independence, Bayes Rules
- Normal Distribution and CLT
- Inference and Hypothesis testing
  - Motivation, Pitfalls and Examples
  - Null/Alternate Hypothesis
  - Test Statistic
  - Sampling Distribution



Significance level

- s = mean(a) – mean(b) is our test statistic, H_0 the hypothesis that mean(A)=mean(B)
  - We reject if $Pr(x > s \mid H_0) < \alpha$
  - α is a suitable "small" probability, say 0.05.

- This threshold probability is the significance level
  - α directly controls the false positive rate (rate at which we expect to observe large s even if is H_0 true).
  - As we make α smaller, the false negative rate increase – situations where mean(A), mean(B) differ but the test fails.
  - Common values 0.05, 0.02, 0.01, 0.005, 0.001

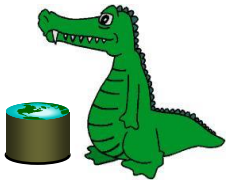


Type I & II Error

type I **error** is the incorrect rejection of a **true** null hypothesis (a "**false positive**")
type II **error** is incorrectly retaining a **false** null hypothesis (a "**false negative**").

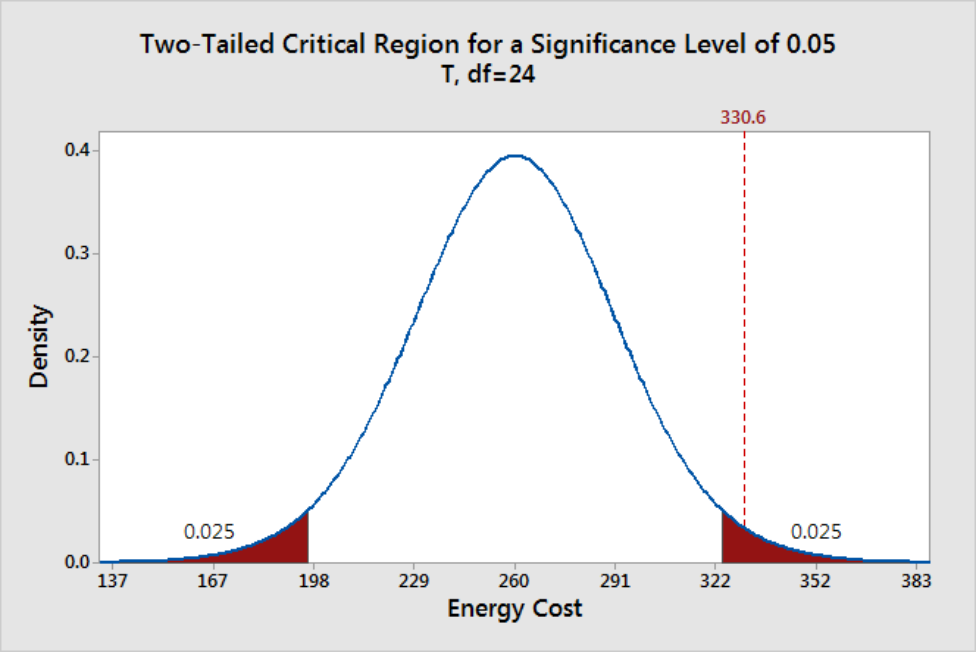
		Decision	
		Retain the null	**Reject the null**
Truth in the population	True	CORRECT $1 - \alpha$	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT $1 - \beta$ POWER

From G.J. Primavera, “Statistics for the Behavioral Sciences”



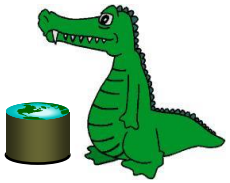
Critical Region I

A significant level α defines a **critical region**, which indicates how far away our sample mean must be from the null hypothesis mean to be statistically significant



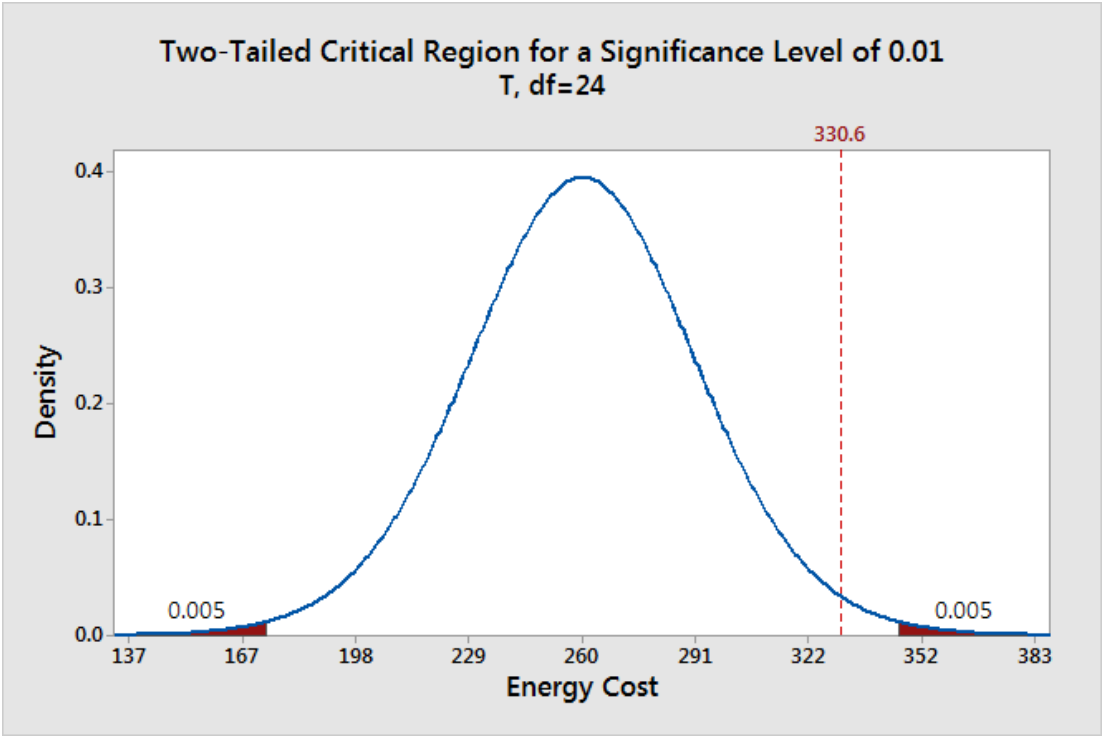
Our sample mean (330.6) falls within the critical region, which indicates it is statistically significant at the $\alpha = .05$ level

*http://blog.minitab.com



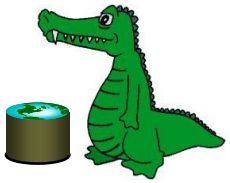
Critical Region II

Another common $\alpha = .01$: our sample mean does not fall within the **critical region**



You needs to choose the significance level before you begin your study: Common α-values are .01, .05, etc.

*http://blog.minitab.com



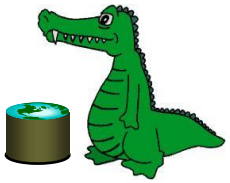
What are p-values

p-value represents the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis.

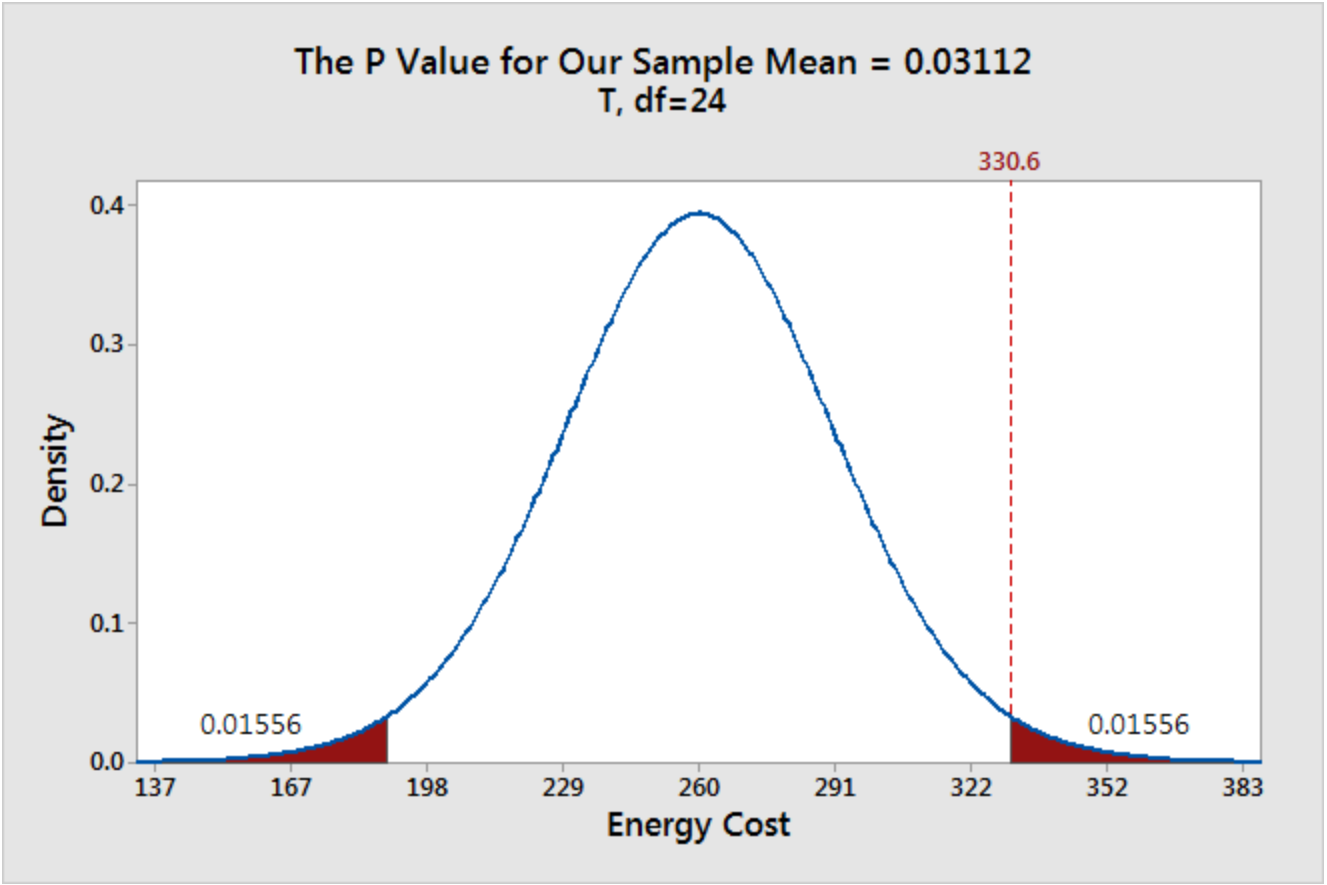
To compute the p-value for our example:

- Compute the distance between our sample mean and the null hypothesis value (330.6 - 260 = 70.6).
- Graph the probability of obtaining a sample mean that is at least as extreme in both tails of the null hypothesis distribution (i.e. 260 +/- 70.6).

*http://blog.minitab.com

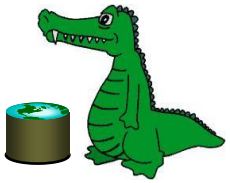


What are p-values



If p-value $\leq$ α, we reject the null hypothesis.

*http://blog.minitab.com

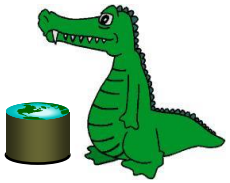


What are p-values

They measure how compatible your data are with the null hypothesis.

- High p-values: data are likely aligned with H_0
- Low p-values: data are unlikely aligned with H_0

But, they don't measure support for the alternative hypothesis.

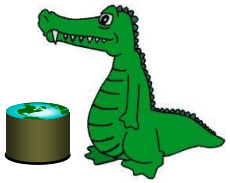


Power

Power of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true.

		Decision	
		Retain the null	**Reject the null**
Truth in the population	True	CORRECT $1 - \alpha$	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT $1 - \beta$ POWER

From G.J. Primavera, “Statistics for the Behavioral Sciences”



One-Sample T-test

Checks the evidence that the mean (μ) of a sample $\neq \mu_0$.

Typical **null hypothesis** and **alternate hypothesis** are:

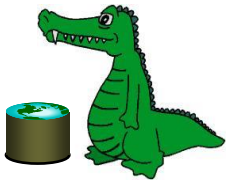
$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

One can also use this test for **one group** of individuals in **two conditions** (before vs. after): use the difference of the two measurements for each person

In this test, we compute the **test statistic** as

$$t = \frac{\bar{X}}{\bar{\sigma}}$$

where $\bar{X}$ is the sample mean and $\bar{\sigma}$ is the sample standard deviation.



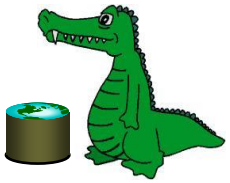
Two sample T-test

Used to compare the means from exactly **TWO** groups, such as the **control** group vs. the **experimental** group

$$H_0\colon \mu(X_1) = \mu(X_2)$$
$$H_a\colon \mu(X_1) \neq \mu(X_2)$$

Suppose there are **two samples** X_1 and X_2. A t-statistic is constructed from their sample means and sample standard deviations:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} \qquad s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$



A/B Testing*

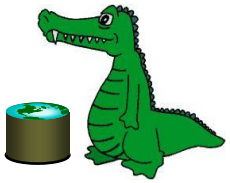
An example problem:

- Test a website landing page that has a signup form
- Test various layouts to try and maximize the "conversion rate", i.e., the percentage of people who sign up

Setup: a **control** treatment and 3 **experimental** treatments A, B, C

Note: It's similar to paired t-test

*http://20bits.com/article/statistical-analysis-and-ab-testing



Synthetic data

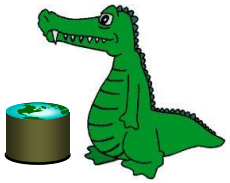
Goal: increase the landing page conversion rate by at least 20%

Project X Landing Page

Treatment	Visitors Treated	Visitors Registered	Conversion Rate
Control	182	35	19.23%
Treatment A	180	45	25.00%
Treatment B	189	28	14.81%
Treatment C	188	61	32.45%

Treatment C is "good enough”, but can you describe the goodness for e.g. with 5% significance level / 95% confidence?

*http://20bits.com/article/statistical-analysis-and-ab-testing



The Statistics

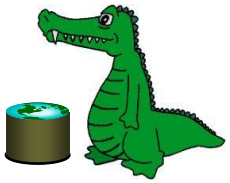
The **null hypothesis**

$$H_0: p - p_c \leq 0$$

p_c is the conversion rate of the control and p is the conversion rate of one of our experiments

The **alternative hypothesis** is that the experimental page has a higher conversion rate.

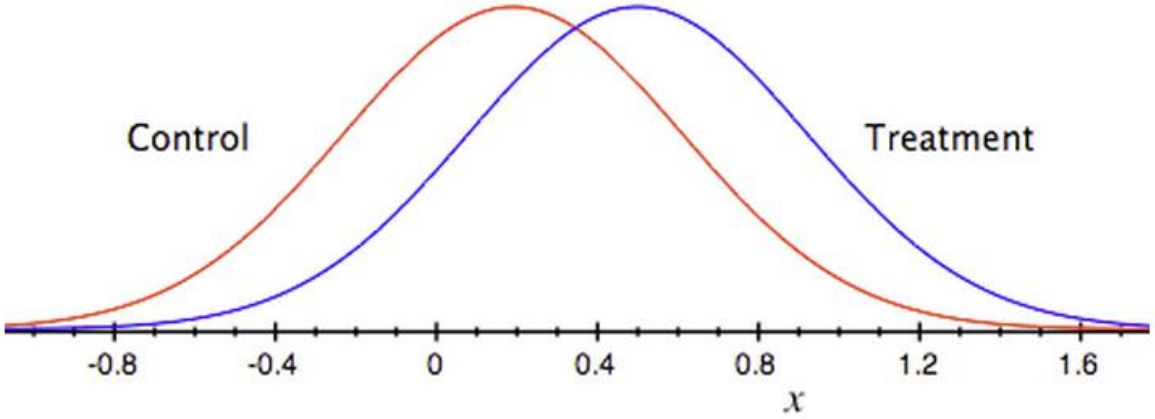
*http://20bits.com/article/statistical-analysis-and-ab-testing



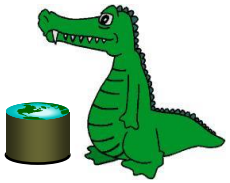
Normality Assumption

The conversion is like a coin flip: heads = "converts" and tails = "doesn't convert"

One can then assume that the sampled conversion rates are normally distributed.



*http://20bits.com/article/statistical-analysis-and-ab-testing



z-scores and One-tailed tests

We define a new r.v. $X = p - p_c$. The null hypothesis becomes

$$H_0: X \leq 0$$

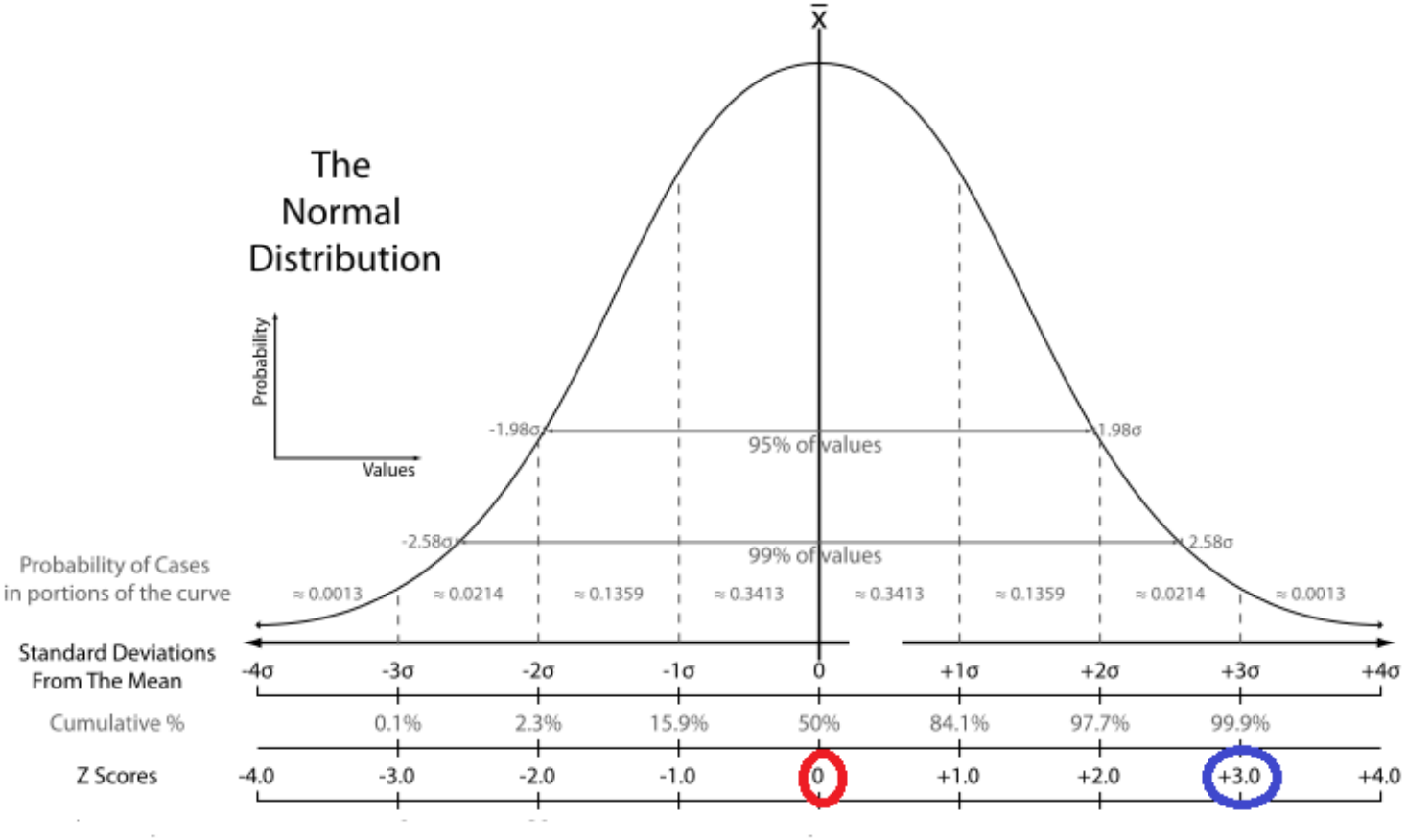
z-score for X: the distance from the population mean to X in terms of standard deviations ($\pm\sigma$)

$$z = \frac{p - p_c}{\sqrt{\frac{p(1-p)}{N} + \frac{p_c(1-p_c)}{N_c}}}$$

We only care about the positive tail of the normal distribution ($H_a: X > 0$)

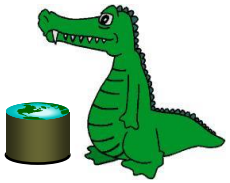
*http://20bits.com/article/statistical-analysis-and-ab-testing

z-scores and One-tailed tests



We reject H_o if the experimental conversion rate is significantly higher (95% confidence i.e. the z-score > 1.65) than the control conversation rate.

* http://www.statisticshowto.com/how-to-calculate-a-z-score/



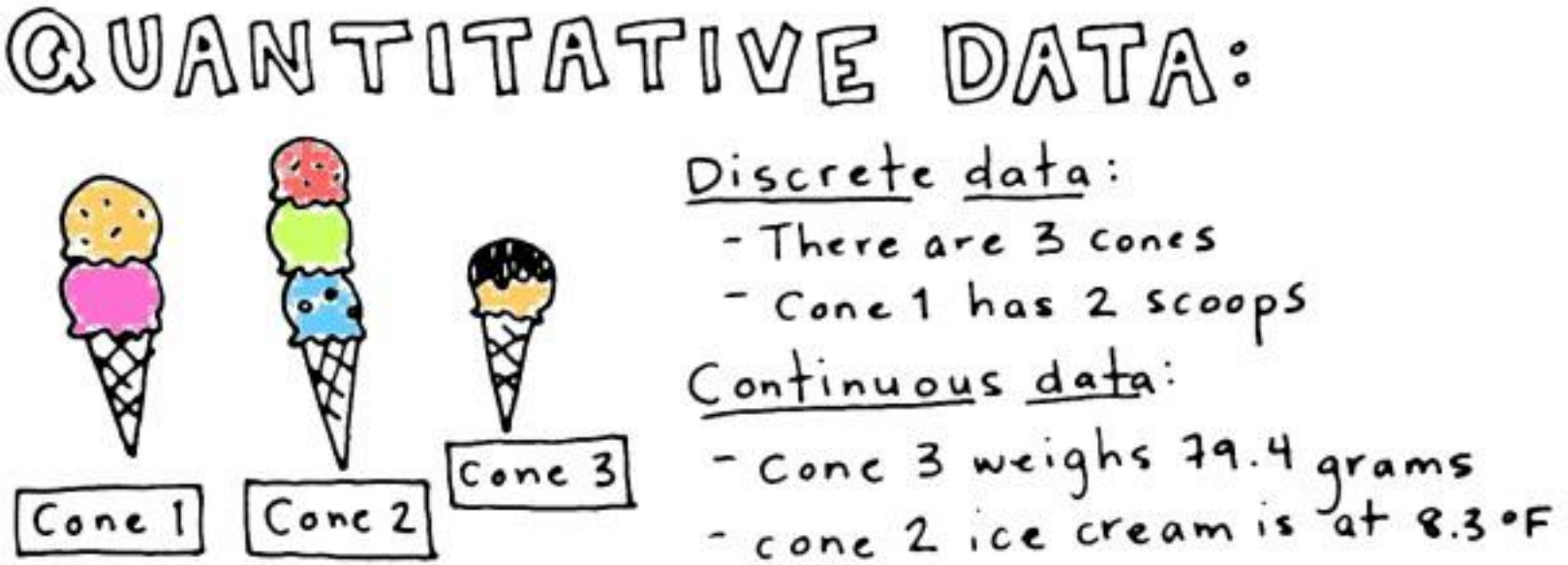
Other important tests

- **T-test:** compare two groups or two interventions on one group.
  - A/B Testing

- **CHI-squared (Fisher's test):** Compare the counts in a "contingency table".

- **ANOVA:** Compare means in more than two different group of individuals.

Data – It's numeric*

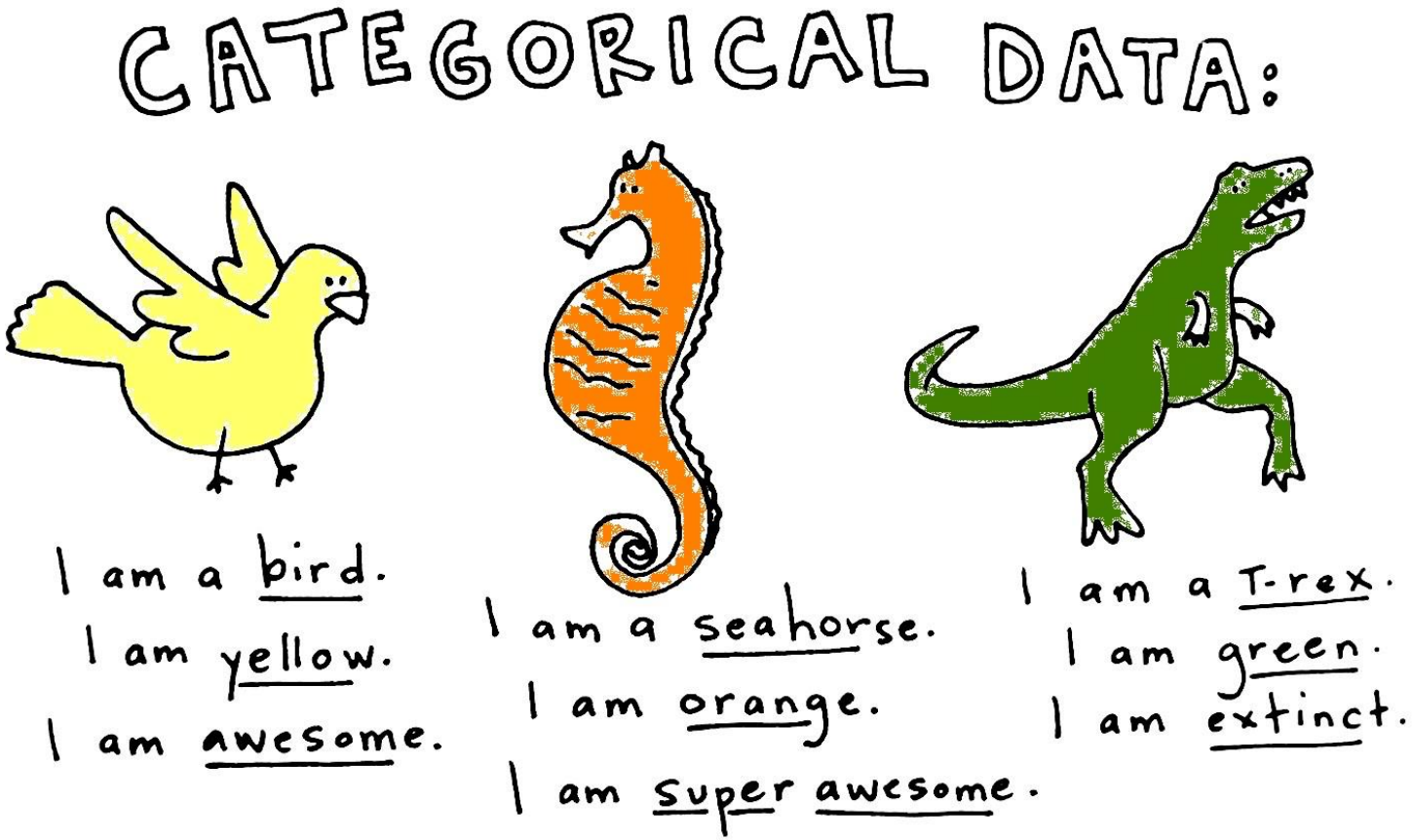


Discrete variables – only a few possible values and no in-between values

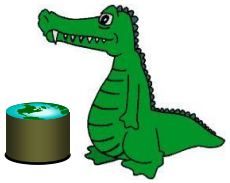
Continuous variables – several possible values and in-between values

*http://statsthewayilikeit.com/

Data – It's descriptive*

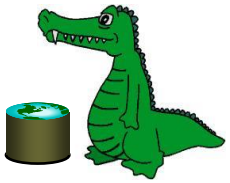


*http://statsthewayilikeit.com/



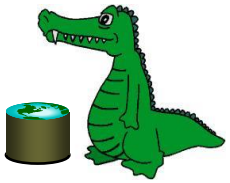
Different types of variables

- A *categorical* variable, also called a *nominal* variable, is for mutual exclusive, but not ordered, categories.
  - E.g. countries
- A *ordinal* variable, is one where the order matters but not the difference between values.
  - E.g, Ranking from search engines
- A *interval* variable is a measurement where the difference between two values is meaningful.
  - E.g., Fahrenheit scale temperature
- A *ratio* variable, has all the properties of an interval variable, and also has a clear definition of 0.0.
  - E.g., height, weight, Kelvin scale temperature



Statistics on different types of variable

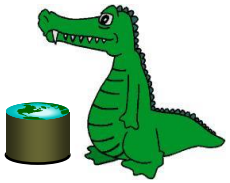
OK to compute....	Nominal	Ordinal	Interval	Ratio
frequency distribution.	Yes	Yes	Yes	Yes
median and percentiles.	No	Yes	Yes	Yes
add or subtract.	No	No	Yes	Yes
mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
ratio, or coefficient of variation.	No	No	No	Yes



Descriptive vs. Inferential Statistics

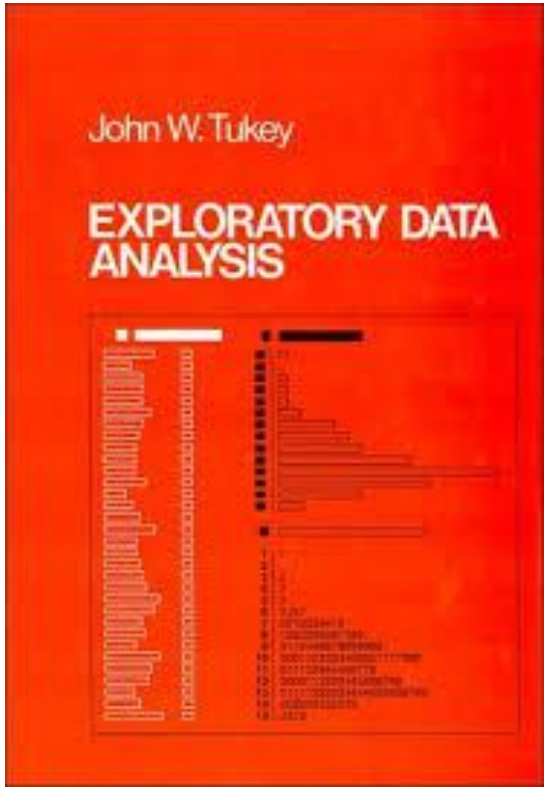
- **Descriptive:** e.g., Median; describes data you have but can't be generalized beyond that

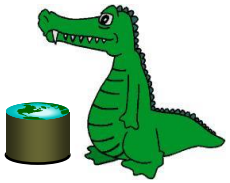
- **Inferential:** e.g., A/B-test, t-test, that enable inferences about the population beyond our data



Start of EDA and statistical computing languages e.g., R

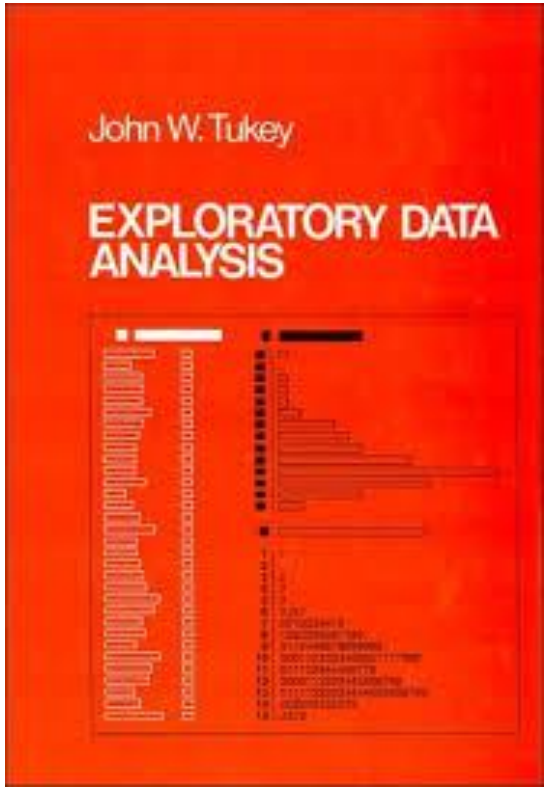
Tukey's championing of EDA encouraged the development of statistical computing packages, especially *S* at Bell Labs. The *S* programming language inspired the systems 'S'-PLUS and *R*. This family of statistical-computing environments featured vastly improved dynamic visualization capabilities, which allowed statisticians to identify outliers, trends and patterns in data that merited further study.

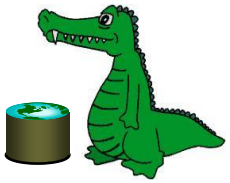




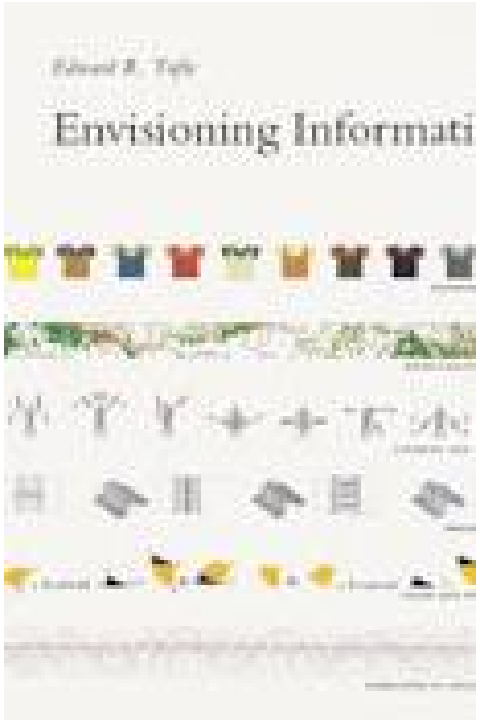
Exploratory Data Analysis, John Tukey 1977

- 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



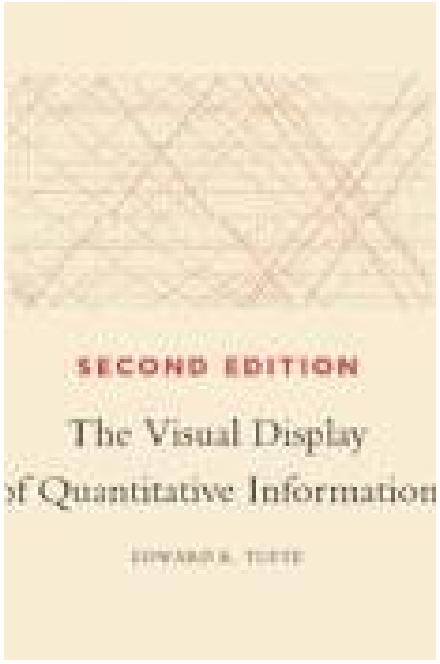


Edward Tufte

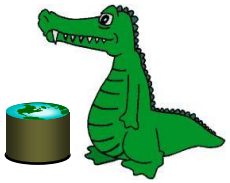


Envisioning Information, 2001

Visual and Statistical Thinking: Displays of Evidence for Making Decisions, 1997



The Visual Display of Quantitative Information, 1983



The Trouble with Summary Stats

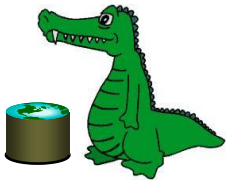
Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15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

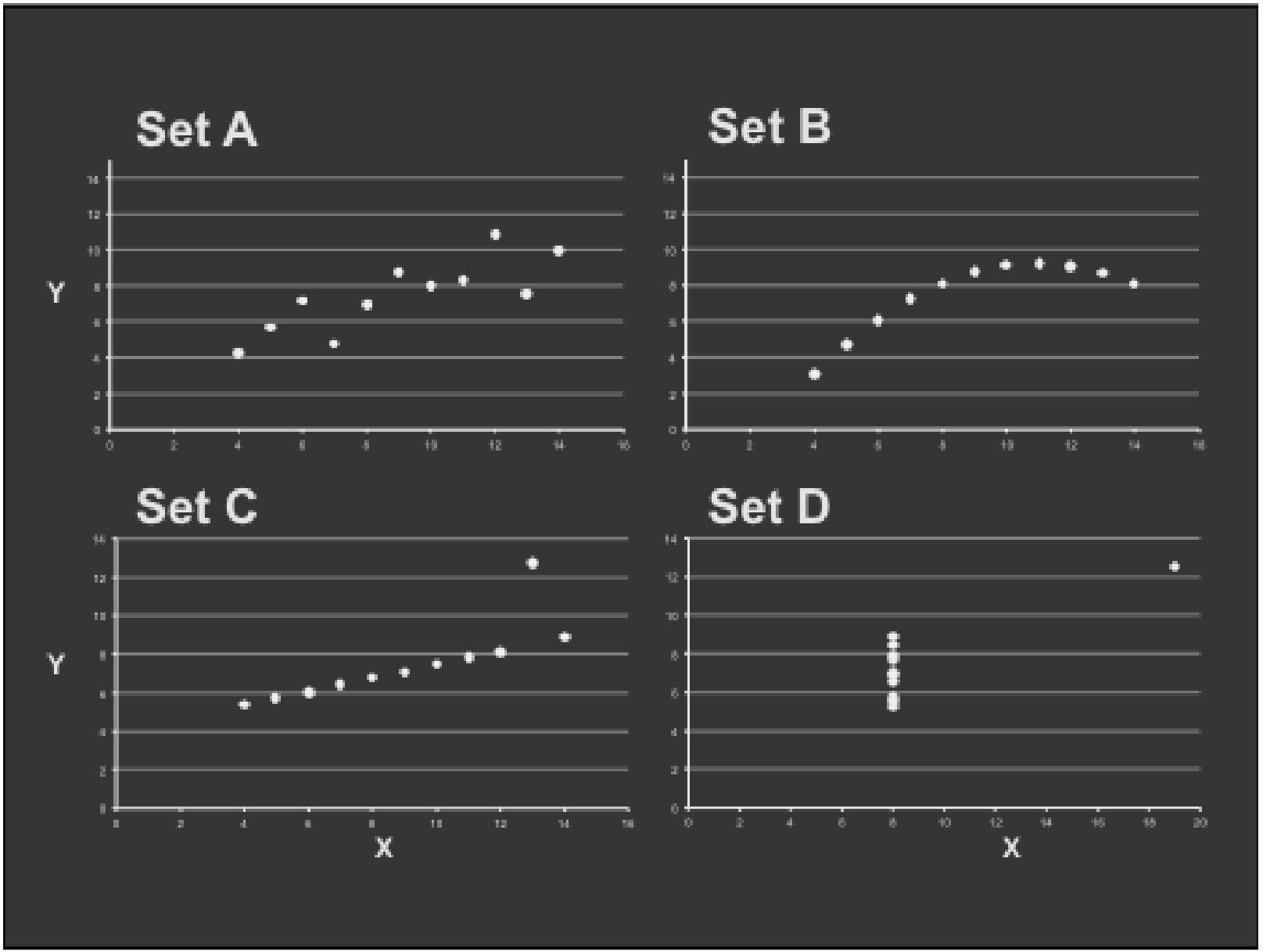
$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\,X$

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

[Anscombe 73]



Looking at Data



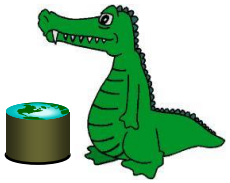
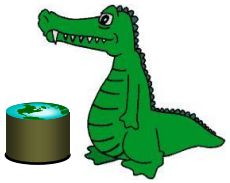


Chart types

- Single variable
  - Dot plot
  - Jitter plot
  - Box-and-whisker plot
  - Histogram
  - Kernel density estimate
  - Cumulative distribution function

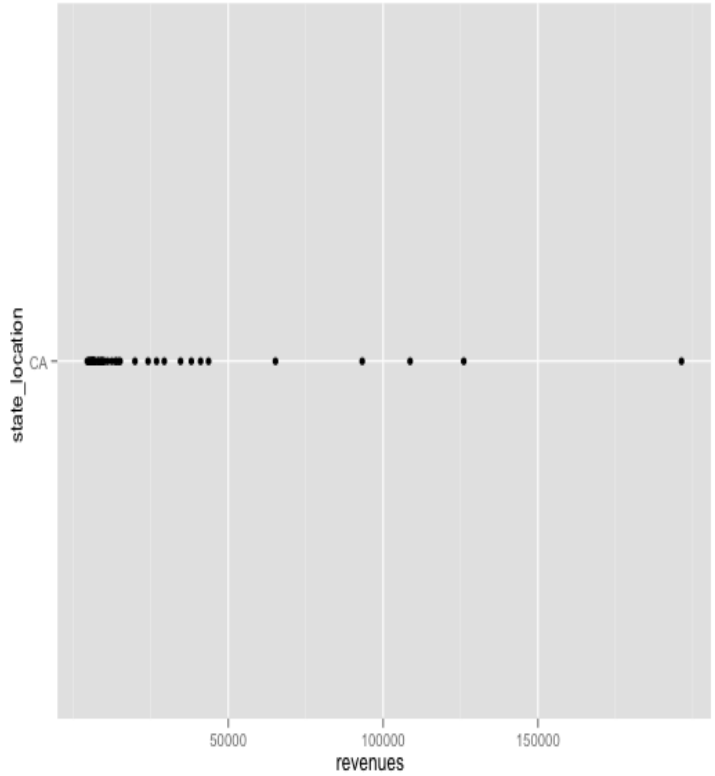
(note: examples using qplot library from R,
But you will be doing it using python)

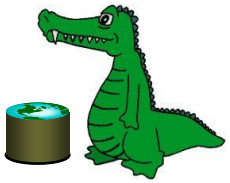
Chart examples from Jeff Hammerbacher's 2012 CS194 class



Dot plot

```
> f500.ca <- subset(f500, state_location == "CA")
> f500.ca$state_location <- factor(f500.ca$state_location)
> qplot(revenues, state_location, data=f500.ca)
```

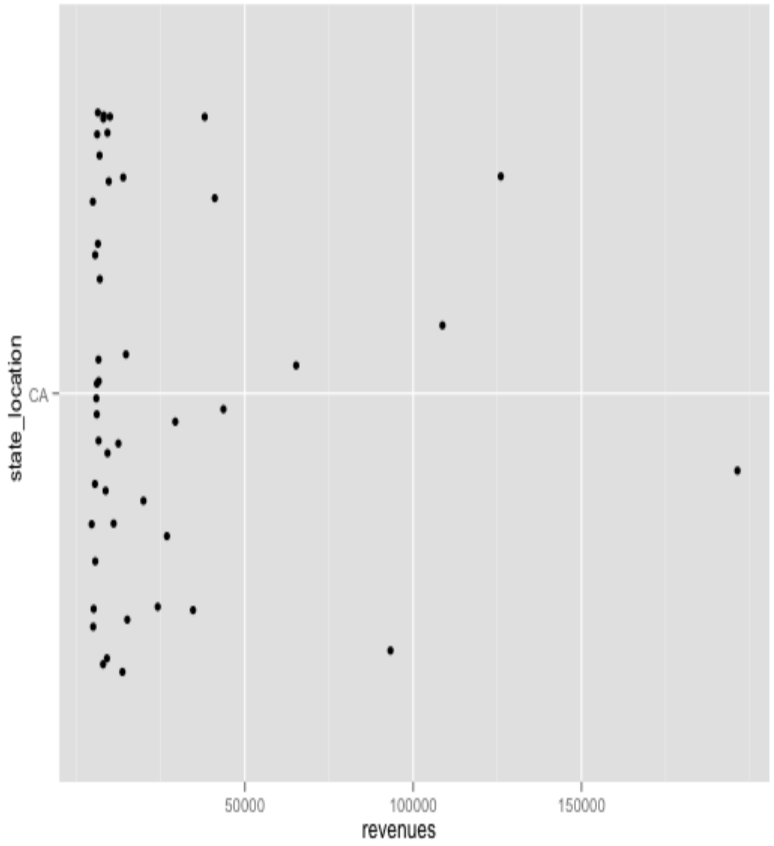


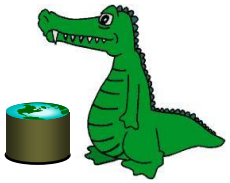


Jitter plot

- Noise added to the y-axis to spread the points

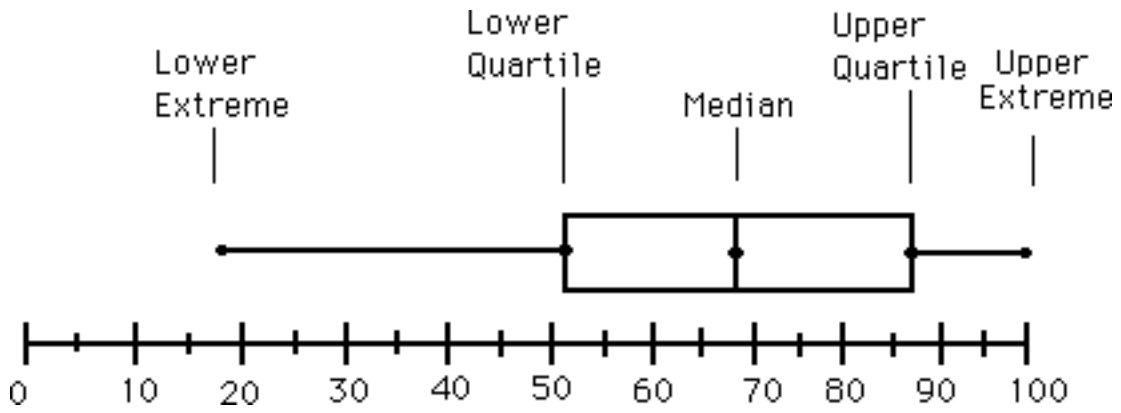
```
> qplot(revenues, state_location, data=f500.ca, geom="jitter")
```

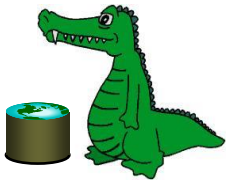




Box-and-whisker plot

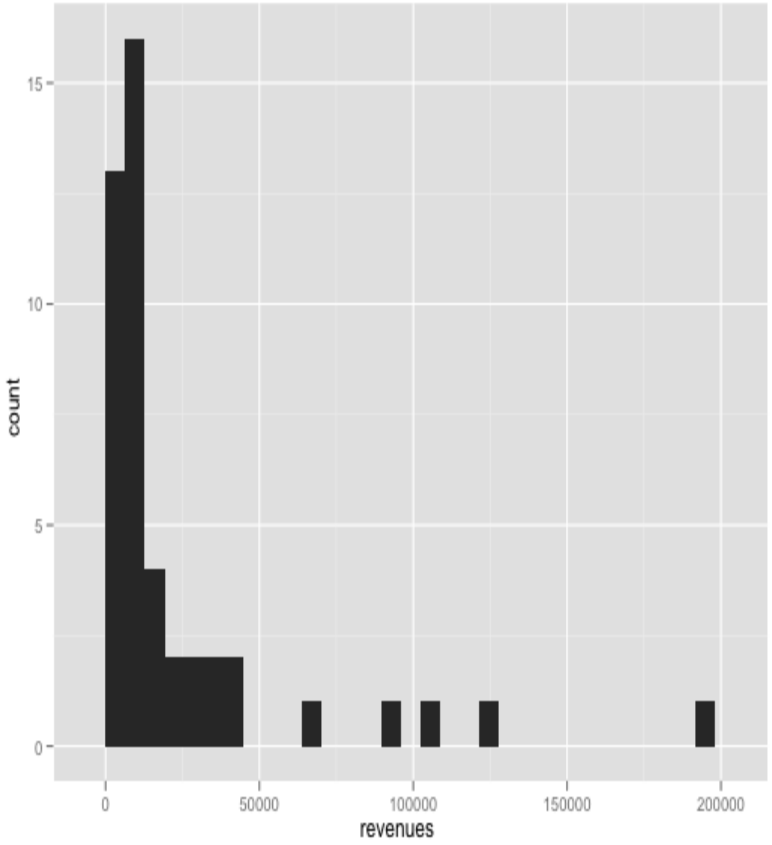
- A graphical form of 5-number summary (Tukey)

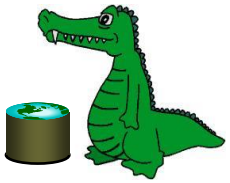




Histogram

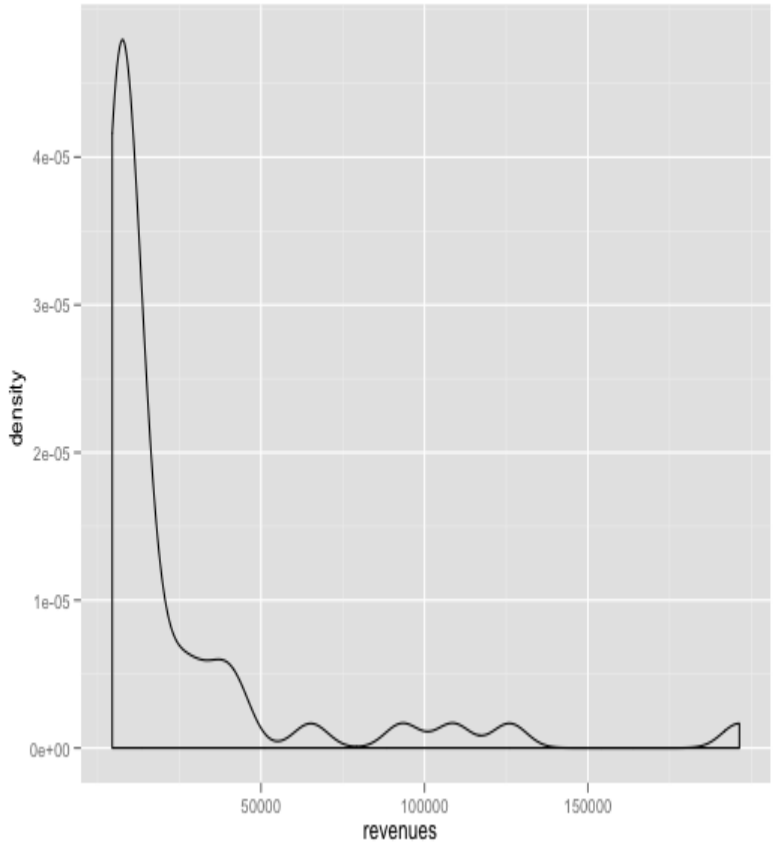
```
> qplot(revenues, data=f500.ca, geom="histogram")
stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

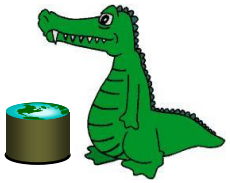




Kernel density estimation: smooth histogram

```
> qplot(revenues, data=f500.ca, geom="density")
```

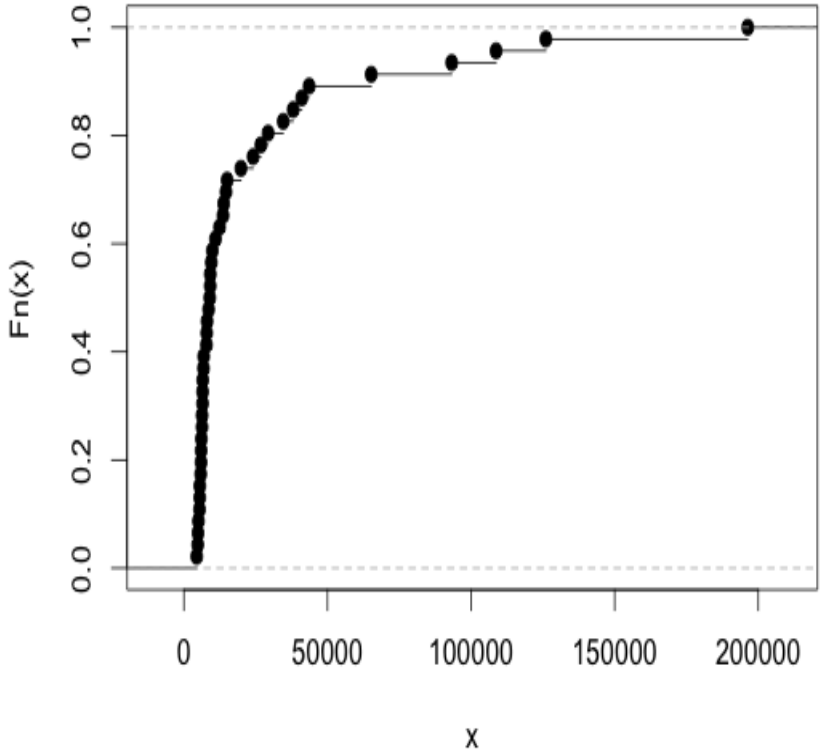


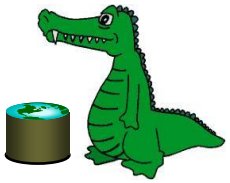


Cumulative distribution function

```
> plot(ecdf(f500.ca$revenues))
```

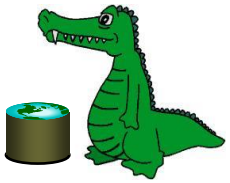
- Integral of the histogram – simpler to build than KDE (don’t need smoothing)





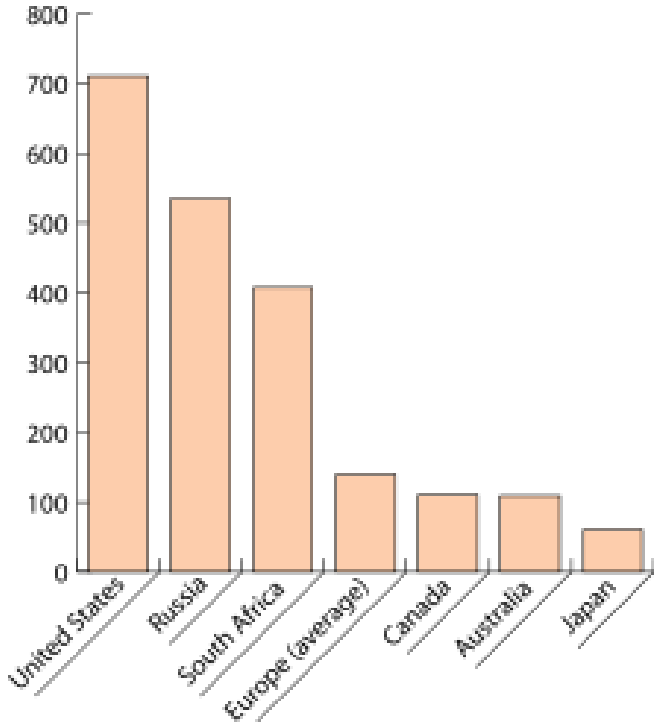
Two-variable Chart types

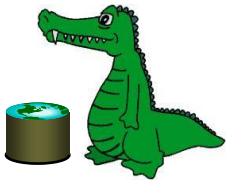
- Bar chart
- Scatter plot
- Line plot
- Log-log plot



Bar plot

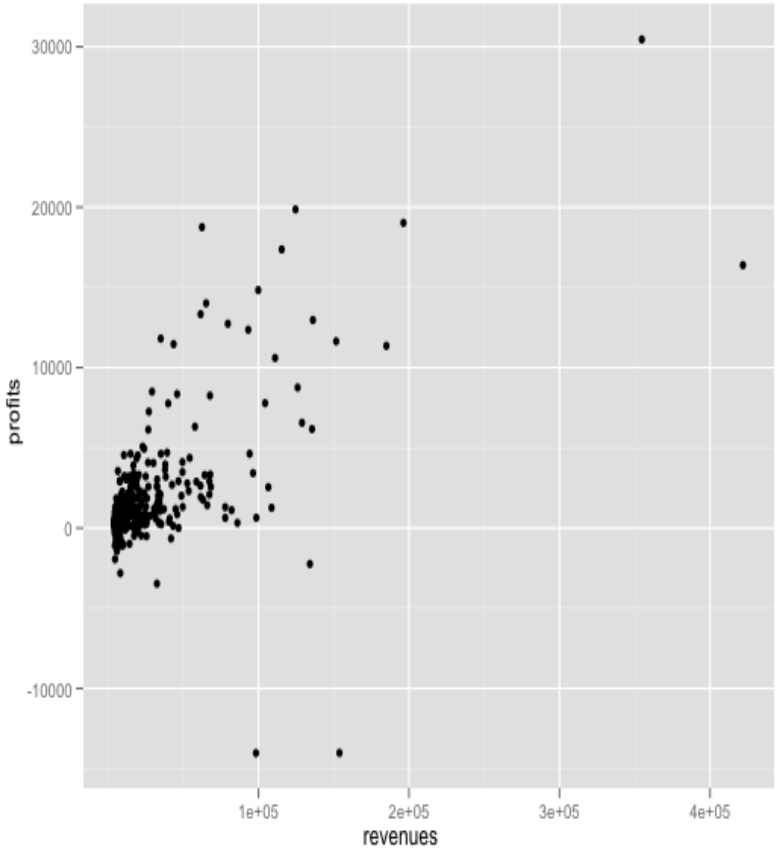
- one variable is discrete

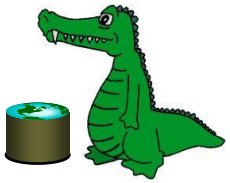




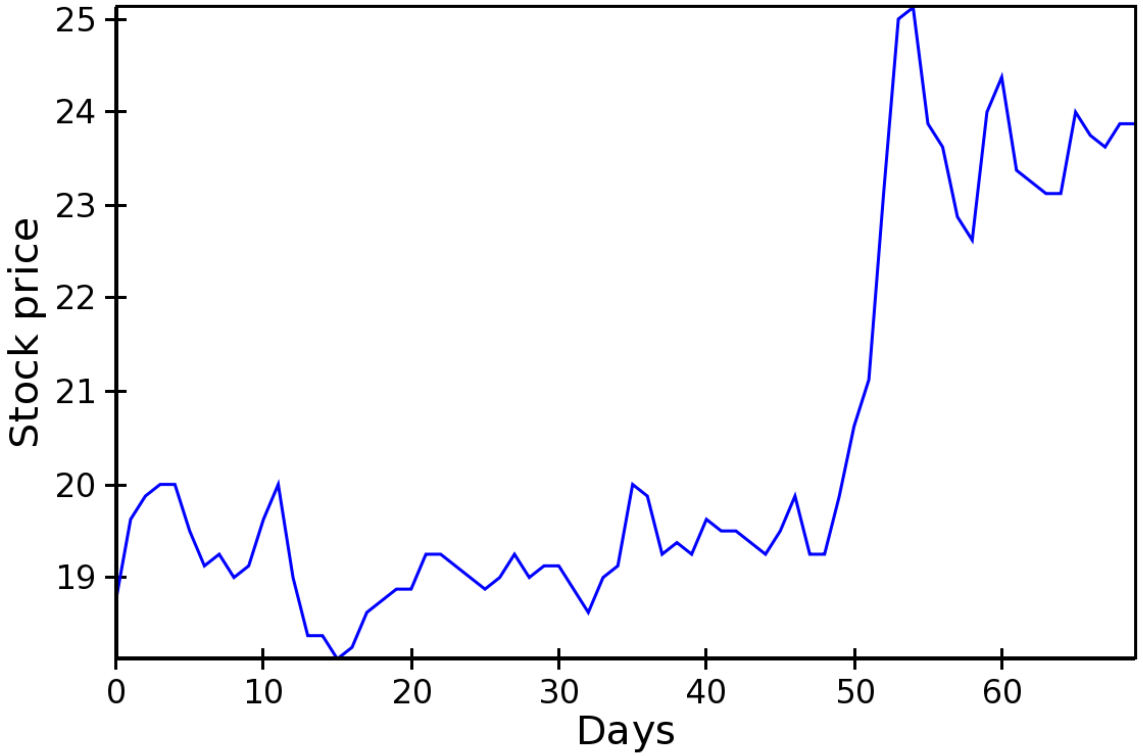
Scatter plot

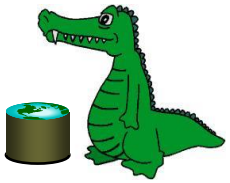
```
> qplot(revenues, profits, data=f500)
```





Line plot

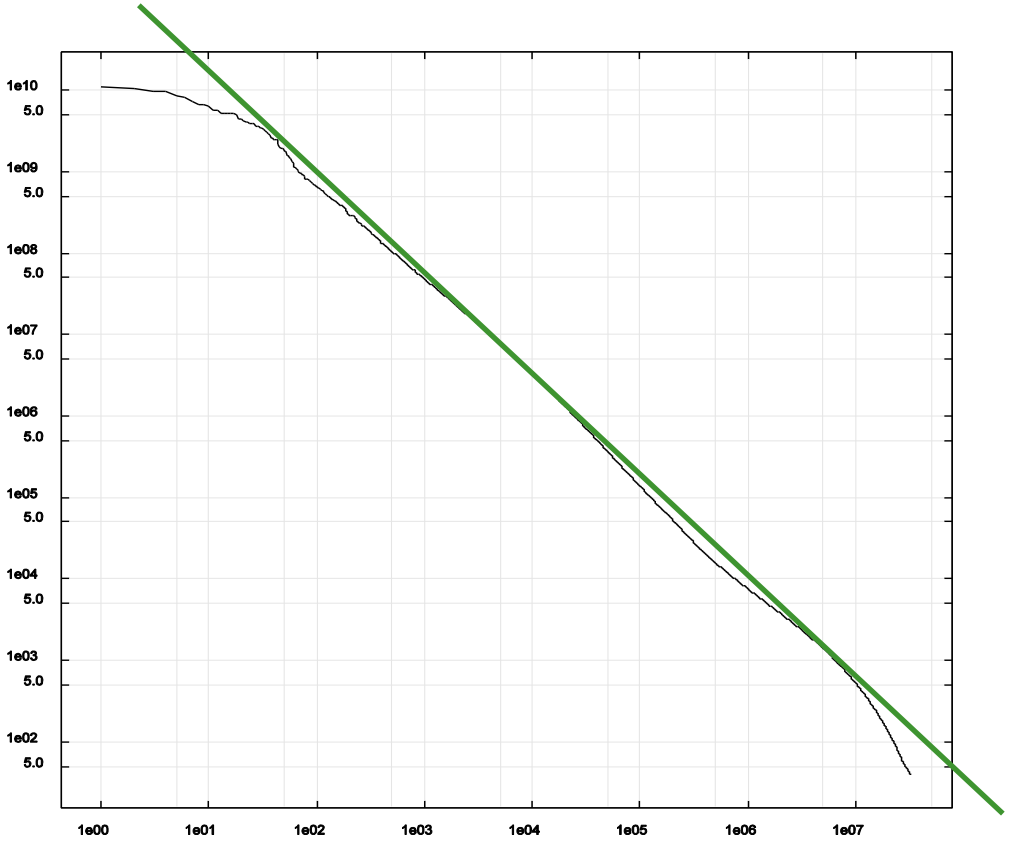




Log-log plot

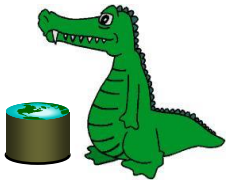
- Very useful for power law data

Frequency of words in tweets



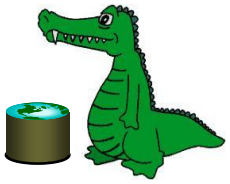
slope ~ -1

Rank of words in tweets, most frequent to least:
I, the, you,…



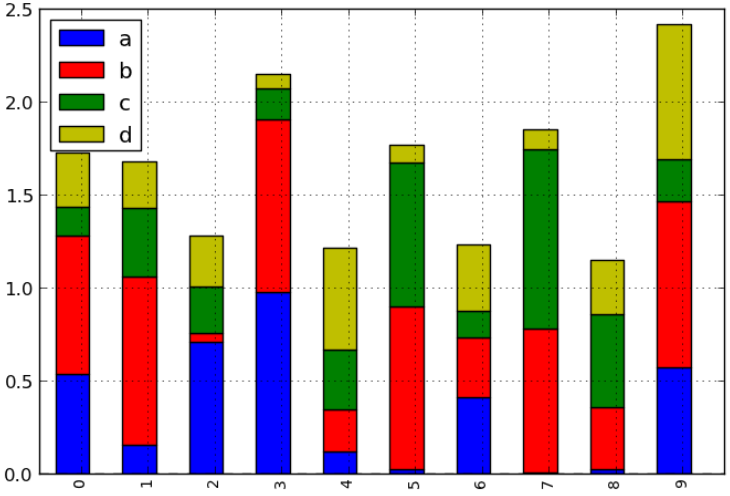
More than two variables

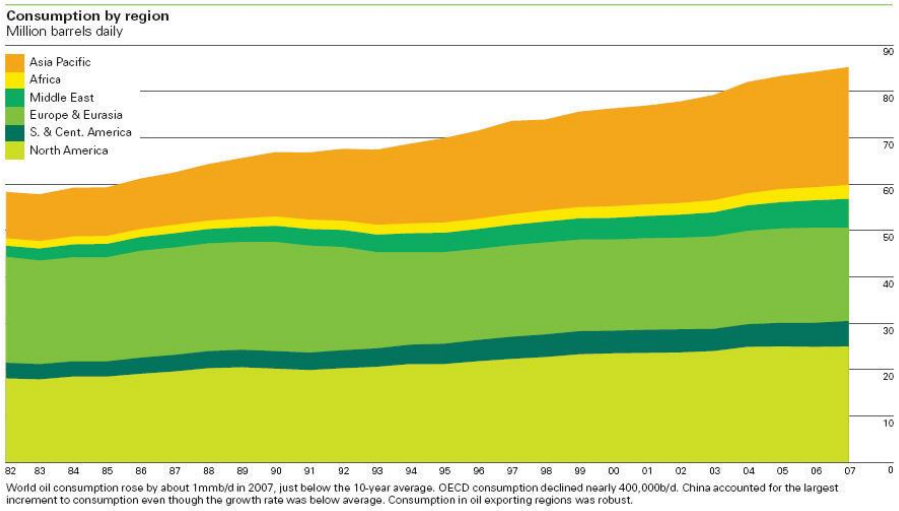
- Stacked plots
- Parallel coordinate plot

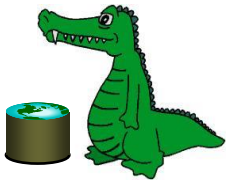


Stacked plot

- stack variable is discrete:

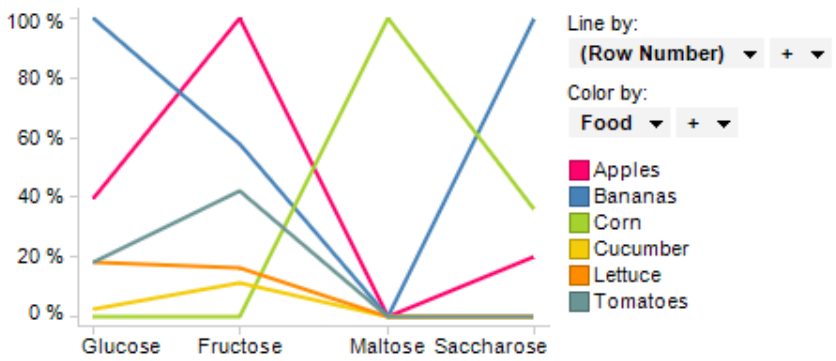


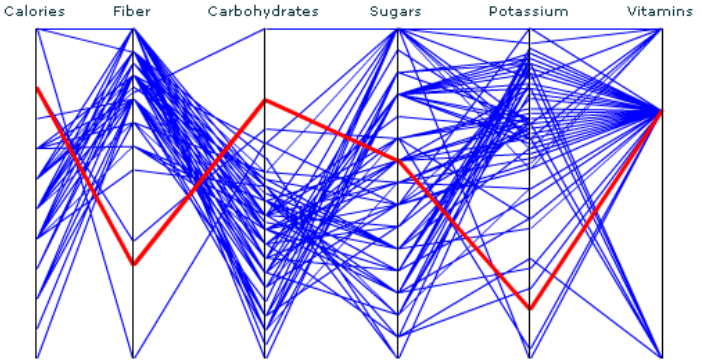


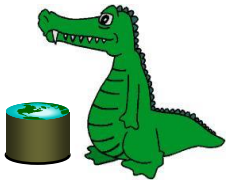


Parallel coordinate plot

- one discrete variable, an arbitrary number of other variables:



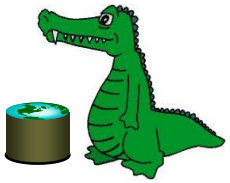




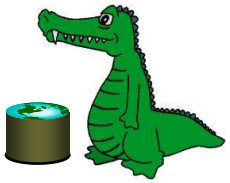
Tools for generating charts

- Python matplotlib, seaborn, etc.
- R qplot

- Tableau
  http://www.tableau.com/learn/whitepapers/which-chart-or-graph-is-right-for-you



Data Presentation can be Art



Summary

- Exploratory Data Analysis
  - Basic statistics
  - Some important distributions
  - Hypothesis Testing
  - Chart types