36-350 Data Mining

Tom Minka

Day 2

Viewing and summarizing batches of numbers

Questions

- Why should we summarize?
- Can we still compare batches fairly?
- Is there an advantage in changing units? Other transformations?

Strip charts

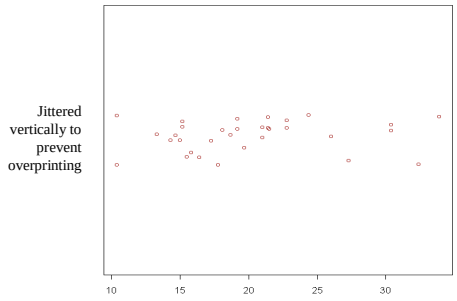
For an unordered batch:

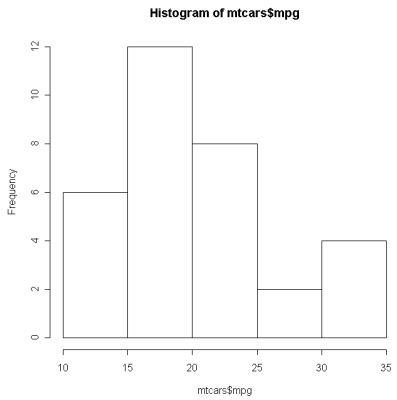
1. Sort
2. Space according to size

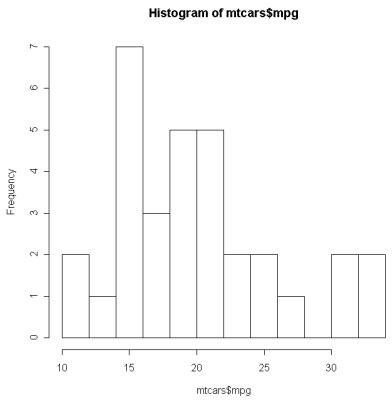
Result is a strip chart

Same principles apply to all graphs

Gas mileage of 32 cars



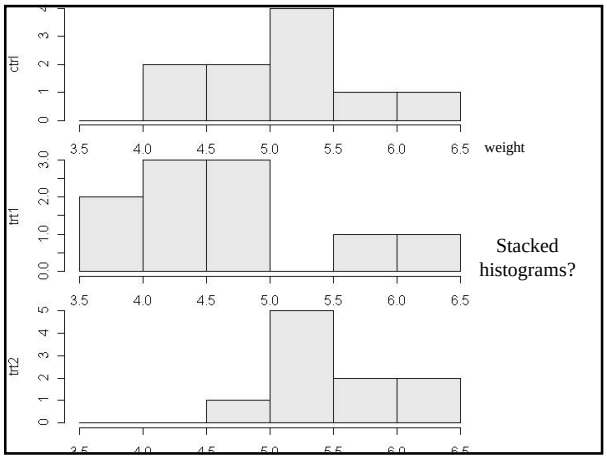


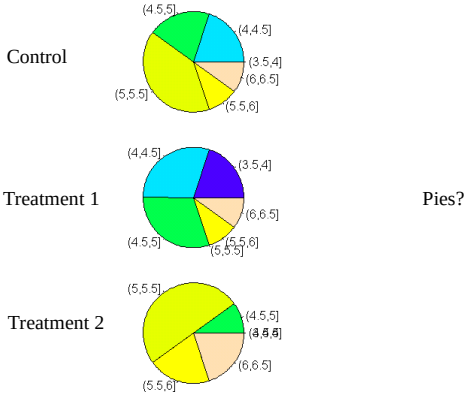


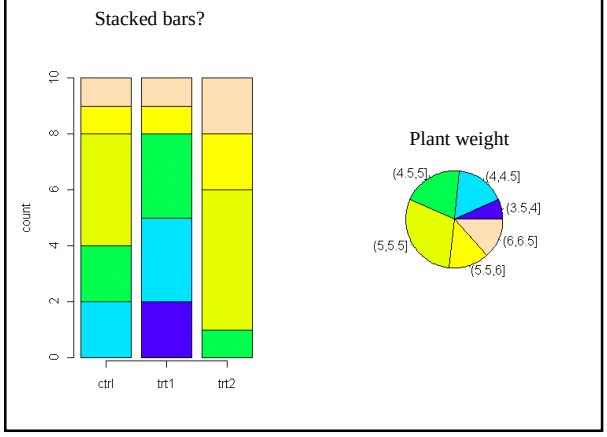
Comparing batches example

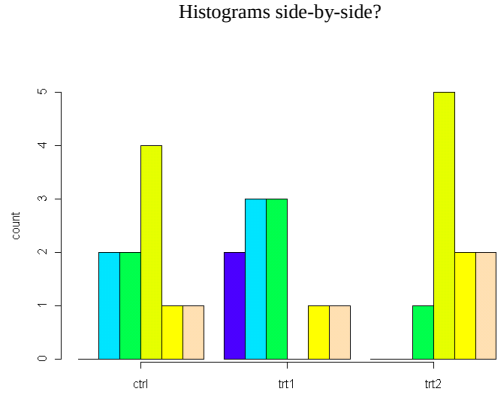
Growing bigger plants

- Three groups of plants:
  - 10 had no treatment (control)
  - 10 had treatment 1
  - 10 had treatment 2
- Compare weights of plants









Defocusing

- Sometimes histograms are too complex
- Can we summarize further?
- Reduce to center, spread, skewness

Estimating "center"

- Mean $\bar{x} = \frac{1}{n}\sum_i x_i$ Minimizes $\sum_i (x_i - \bar{x})^2$
- Median Middle number in sorted order

  Minimizes $\sum_i |x_i - M|$

Mean vs. median

- Mean can be swayed by a single wild number
- Median is "resistant"
- Also depends on underlying source of variation
  - Mean is theoretically better for some sources, median for others

Estimating "spread"

- Standard deviation $s = \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2}$
- Mean abs deviation $MAD = \frac{1}{n}\sum_i |x_i - M|$
- Interquartile range $IQR = Q1 - Q3$

  Q1 = 25th percentile

  M = 50th percentile

  Q3 = 75th percentile

Std dev vs. IQR

- IQR is resistant, standard dev is not
- Quartiles can measure symmetry

$\frac{Q1+Q3}{2} \approx M$ symmetric distribution

$\frac{Q1+Q3}{2} < M$ skewed to **left**

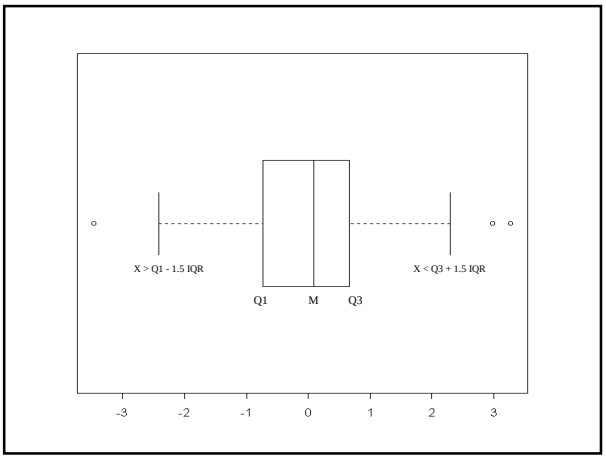
$\frac{Q1+Q3}{2} > M$ skewed to **right**

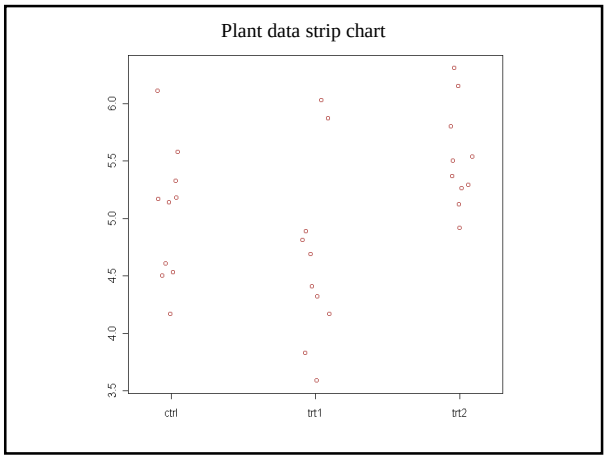
Outsiders

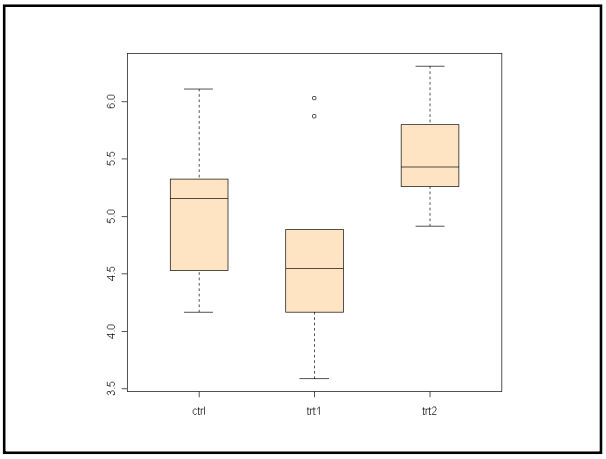
- Fences are $Q1 - 1.5IQR,\ Q3 + 1.5IQR$
- "Outside" points are outside the fences
- Improbable, but not impossible
- Not necessarily corrupted values
- Often the most interesting points

Boxplot

- Box around (Q1,Q3), line at M
- Whiskers to outermost inside points (not fences)
- Outsiders as dots



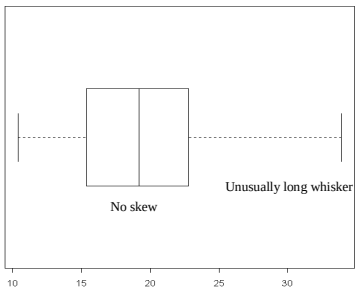




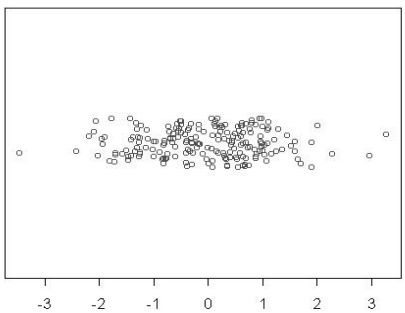
Boxplots

- Boxplots make interpretation easier
- while not hiding too much
- Cases where boxplots fail:
  - Distributions with multiple peaks
  - High amounts of skew

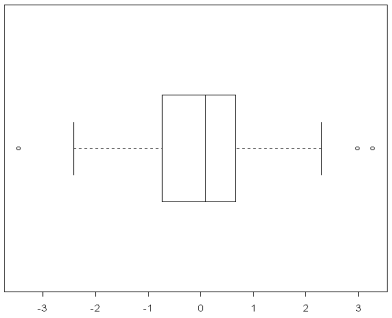
Boxplot of mileage data



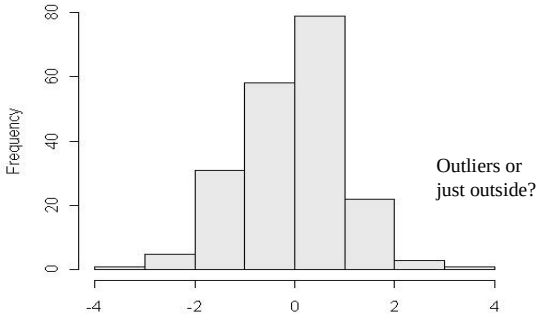
Strip chart of 200 points



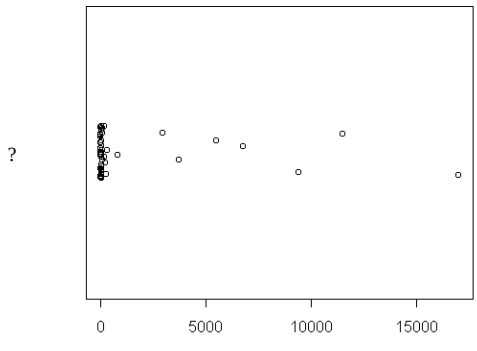
Boxplot of the batch

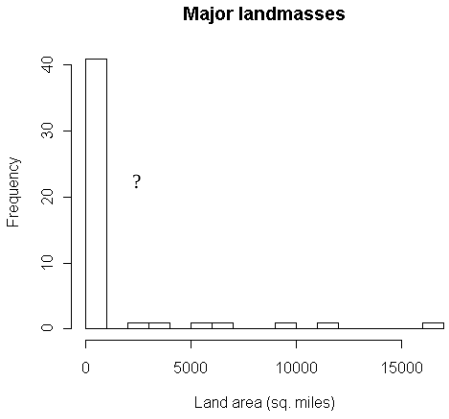


Histogram



A highly skewed dataset





Transformation

- Original representation is not always the best one
- Transformation can remove skew, reduce number of “outside” points

Transformation

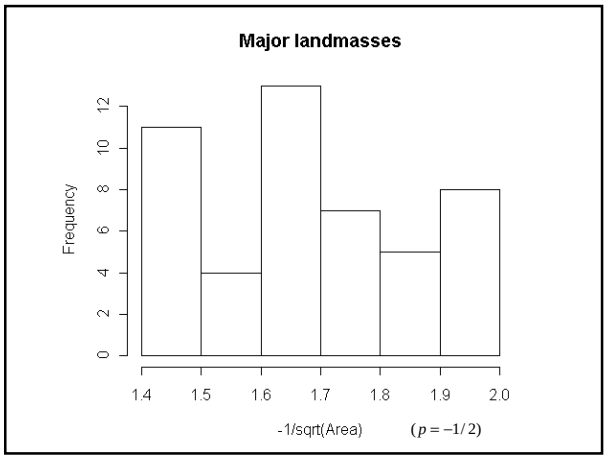
- It doesn't help to change units (e.g. temperature)

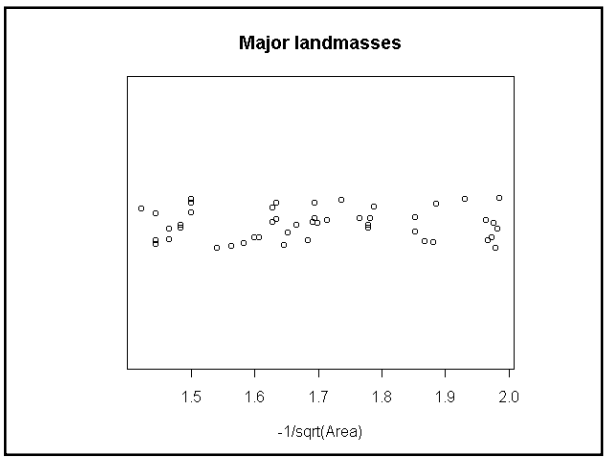
$$x_{new} = ax + b$$

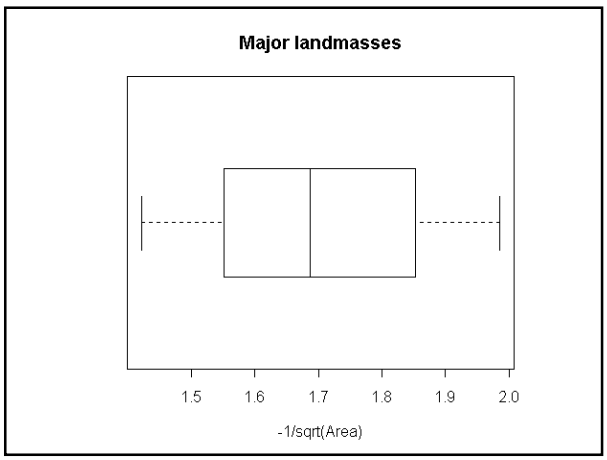
- But nonlinear transformations may help

Roots/reciprocals: $x_{new} = \frac{x^p - 1}{p}$ if $p > 0$ or $p < 0$

Logarithms: $x_{new} = \log(x)$ $\lim p \to 0$







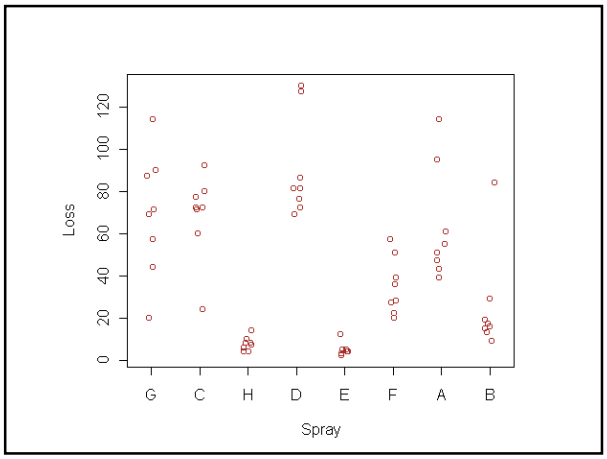
Transformation to remove skew

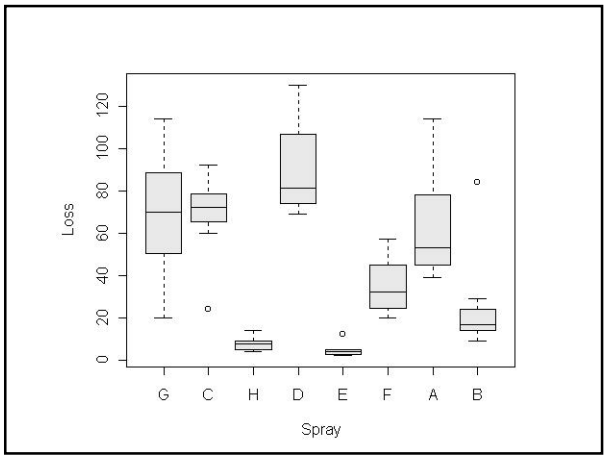
- Reduces outside points, impact of outliers
- Allows sharper description of data
  - "uniform after transformation" vs. "skewed to right"
- Can use standard methods on result
- May stabilize spread of different groups
- Simplifies so we can consider more

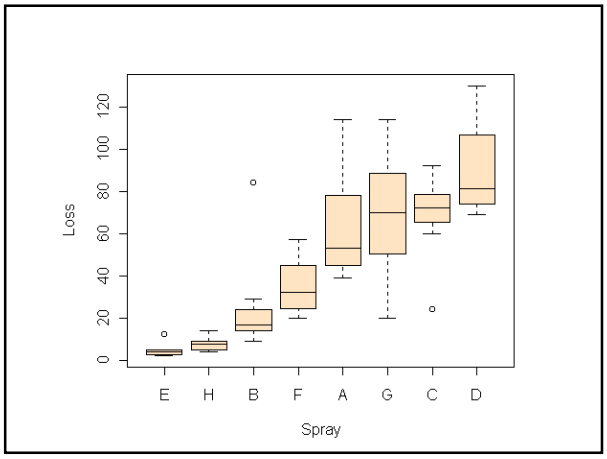
Extended example

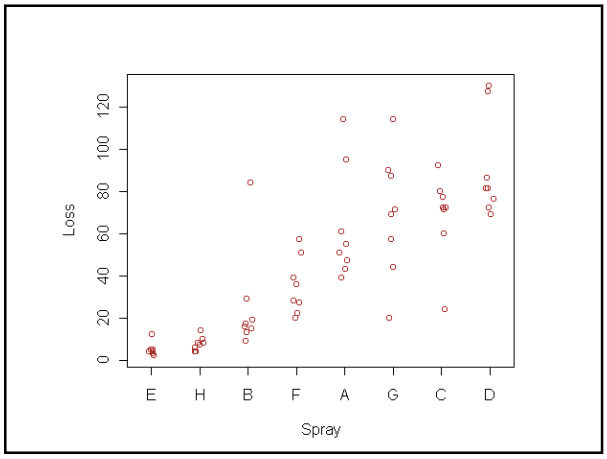
What is the best spray for repelling bees?

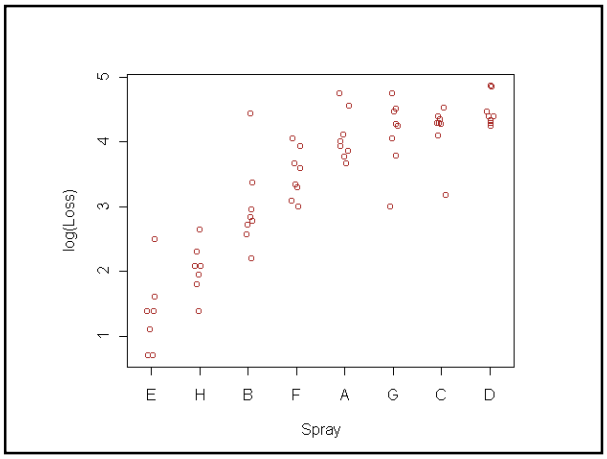
- 8 different sprays
- Each tried 8 different times (8x8 design)
- "loss" of sugar solution reported for each
  - More loss = less repelling

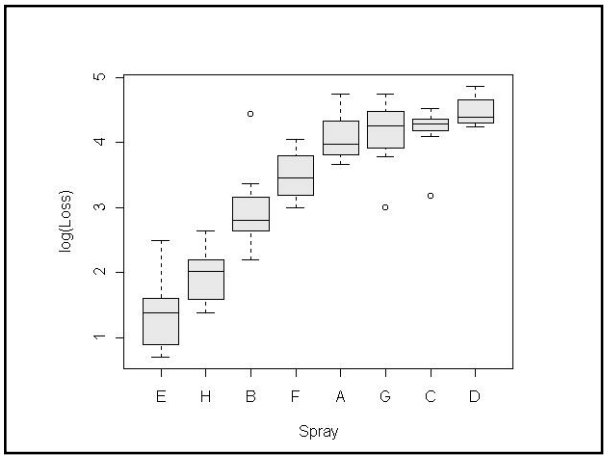

Loss
0
20
40
60
80
100
120
G
C
H
D
E
F
A
B
Spray

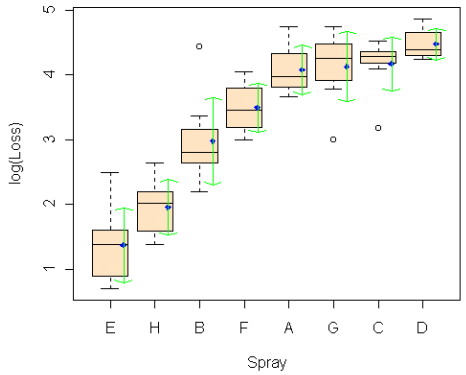

Loss
0
20
40
60
80
100
120
G
C
H
D
E
F
A
B
Spray

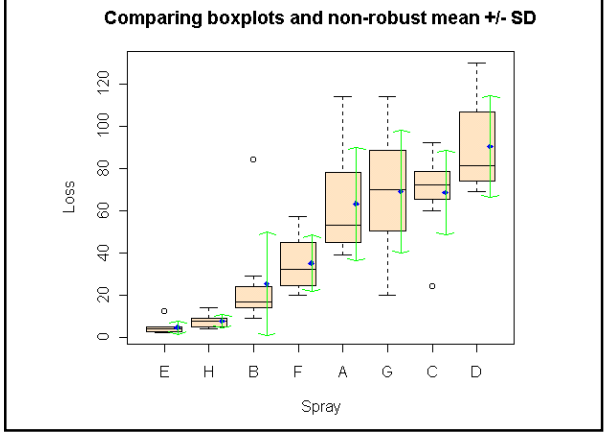

Loss
0
20
40
60
80
100
120
E
H
B
F
A
G
C
D
Spray


Loss
0
20
40
60
80
100
120
E
H
B
F
A
G
C
D
Spray


log(Loss)
1
2
3
4
5
E
H
B
F
A
G
C
D
Spray


log(Loss)
1
2
3
4
5
E
H
B
F
A
G
C
D
Spray



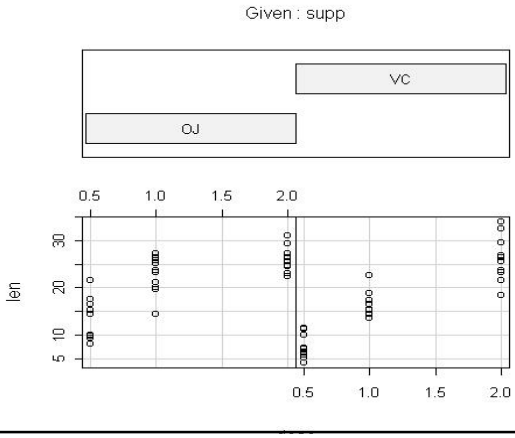


Comparing response across more than one factor

Vitamin C supplements vs. OJ

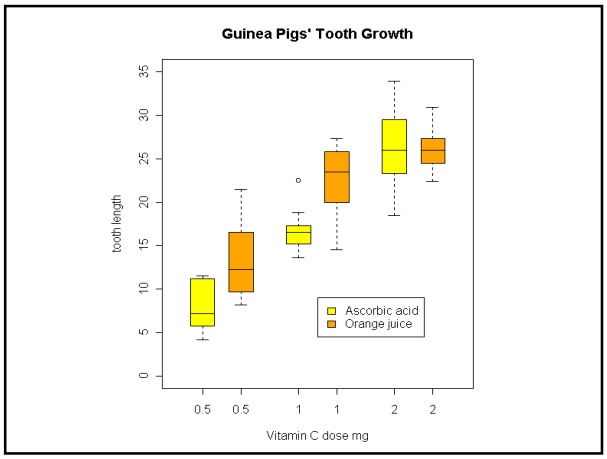
- Guinea pigs were given Vitamin C either in OJ or as ascorbic acid
- Is there a difference in growth (e.g. tooth length)?

Divide by supplement, then dose?



Divide by dose, then supplement

Divide by major, less interesting effects first


Guinea Pigs' Tooth Growth
tooth length
0
5
10
15
20
25
30
35
0.5
0.5
1
1
2
2
Vitamin C dose mg
Ascorbic acid
Orange juice