

Comparative Box Plot Layouts

A box plot provides an easy to grasp distribution caricature

Conveys the location, spread and outliers.

Suitably simple for use in comparisons

Distribution features indexed by factors (categorical variables) can be compared.

1st quartile, 2nd quartile (median), 3rd quartile, adjacent values and outliers

Alternative quartile language: 25th, 50th and 75th percentiles

One page layouts

Small multiples can interest readers and motivate discussion

Edward Tufte says well-designed small multiples are

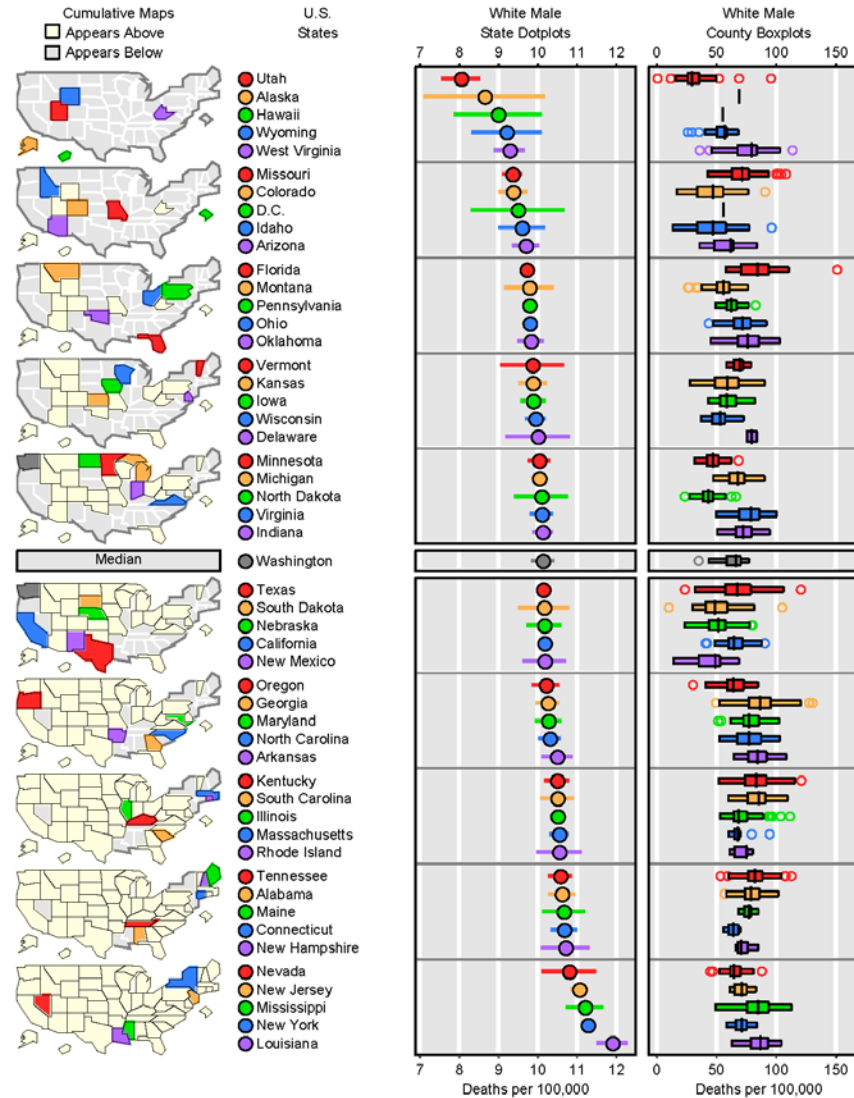
“inevitable comparative, deftly multivariate, ...,
efficient in interpretation and often
narrative in context, and out”

A linked micromap example

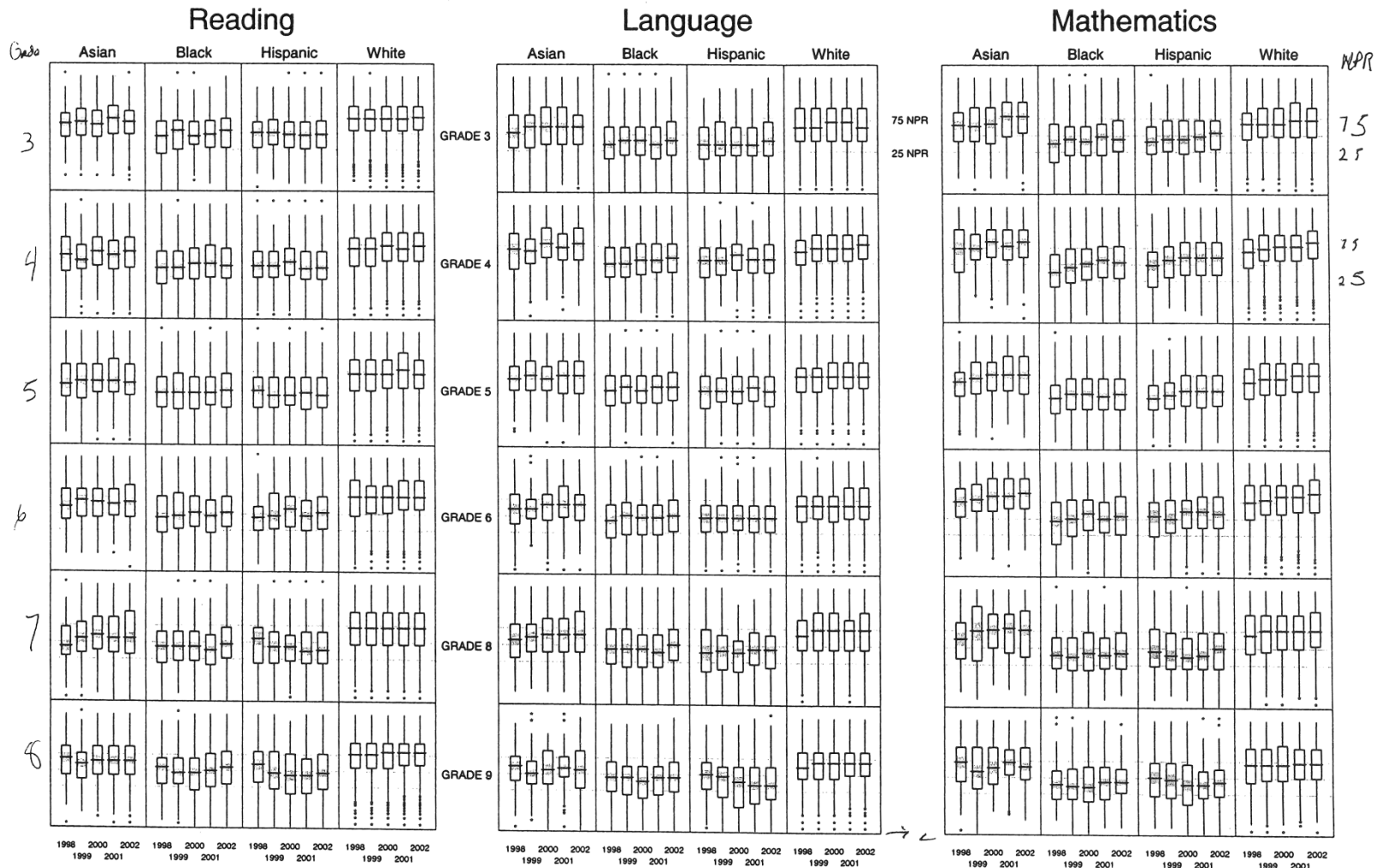
Student examples from past classes

Stanford 9 NCE (Normal Curve Equivalent) scores 2002 for a Virginia county
Percent Change between two versions of Aircraft Emission Models

Lung Cancer for State and County Males

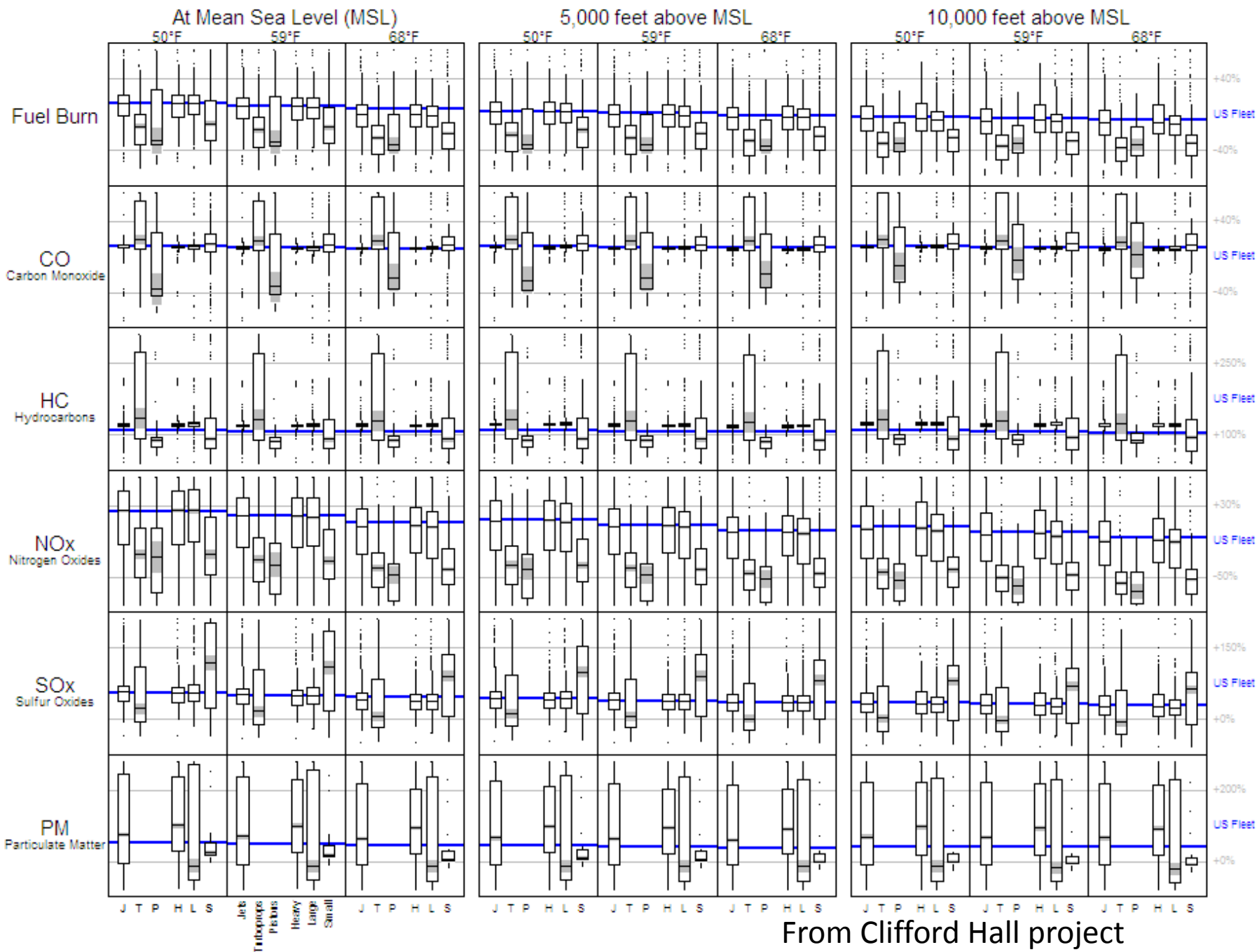


2002 Stanford 9 NCE Scores



Allow's Large Font

Percent Change in Emissions from EDMS 4.5 to 5.0



A Data Mining Visualization Example

Drug and symptom database

Doctors or medical institutions send in reports about individual patients. The reports list drugs and symptoms. Millions of reports are sent in.

Multiple reports can go in for the same patient. If a patient takes an additional or new drug and new symptoms appear, the combination can be of interest. It makes sense to send new reports.

Problems

There are people and times for which drug and symptom combinations go unreported.

Further reports don't include 1) drugs with no symptoms, 2) symptoms with no drugs, and 3) either symptoms or drugs. The absence of this information restricts inference. Denominators are not available for estimating proportions.

A Data Mining Visualization Example

Drug and symptom database

Doctors or medical institutions send in reports about individual patients. The reports list drugs and symptoms. Millions of reports are sent in.

Multiple reports can go in for the same patient. If a patient takes an additional or new drug and new symptoms appear, the combination can be of interest. It makes sense to send new reports.

Problems

There are people and times for which drug and symptom combinations go unreported.

Further reports don't include 1) drugs with no symptoms, 2) symptoms with no drugs, and 3) either symptoms or drugs. The absence of this information restricts inference. Denominators are not available for estimating proportions.

A Data Mining and Visualization Example

Data and Problems

Drug and symptom database

Doctors or medical institutions send in reports about individual patients. The reports list drugs and symptoms. Millions of reports are sent in.

Multiple reports can go in for the same patient. If a patient takes an additional or new drug and new symptoms appear, the combination can be of interest. It makes sense to send new reports.

Problems

There are people and times for which drug and symptom combinations go unreported. There can be bias problems.

Further reports don't include 1) drugs with no symptoms, 2) symptoms with no drugs, and 3) either symptoms or drugs. The absence of this information restricts inference. Denominators are not available for estimating proportions.

A Data Mining and Visualization Example


Analysis Plan and Problems

Plan A

Uses the reported occurrences as a reference. Estimates the likelihood for drug A and Symptom B appearing together using the margin frequencies in table. If the observed number of joint appearances is much higher than expected, then the combination is suspect. The ratio of observed to expected occurrences is labeled the **relative risk, a reasonable statistic to use in making comparisons**.

Problems:

Assuming independence to compute the expected number of occurrences is not necessarily justified. The pooling of information across strata such as organ systems can lead examples exhibiting Simpson's Paradox



A Data Mining and Visualization Example

DuMouchel and Pregibon Approach

Perform a stratified analysis based on Organ Systems

Use an empirical Bayes model that shrinks relative risk estimates toward 1.

Shrinkage is greatest when little data backs up estimates. Impacts high relative risks. Provides a framework for controlling the number false positives.

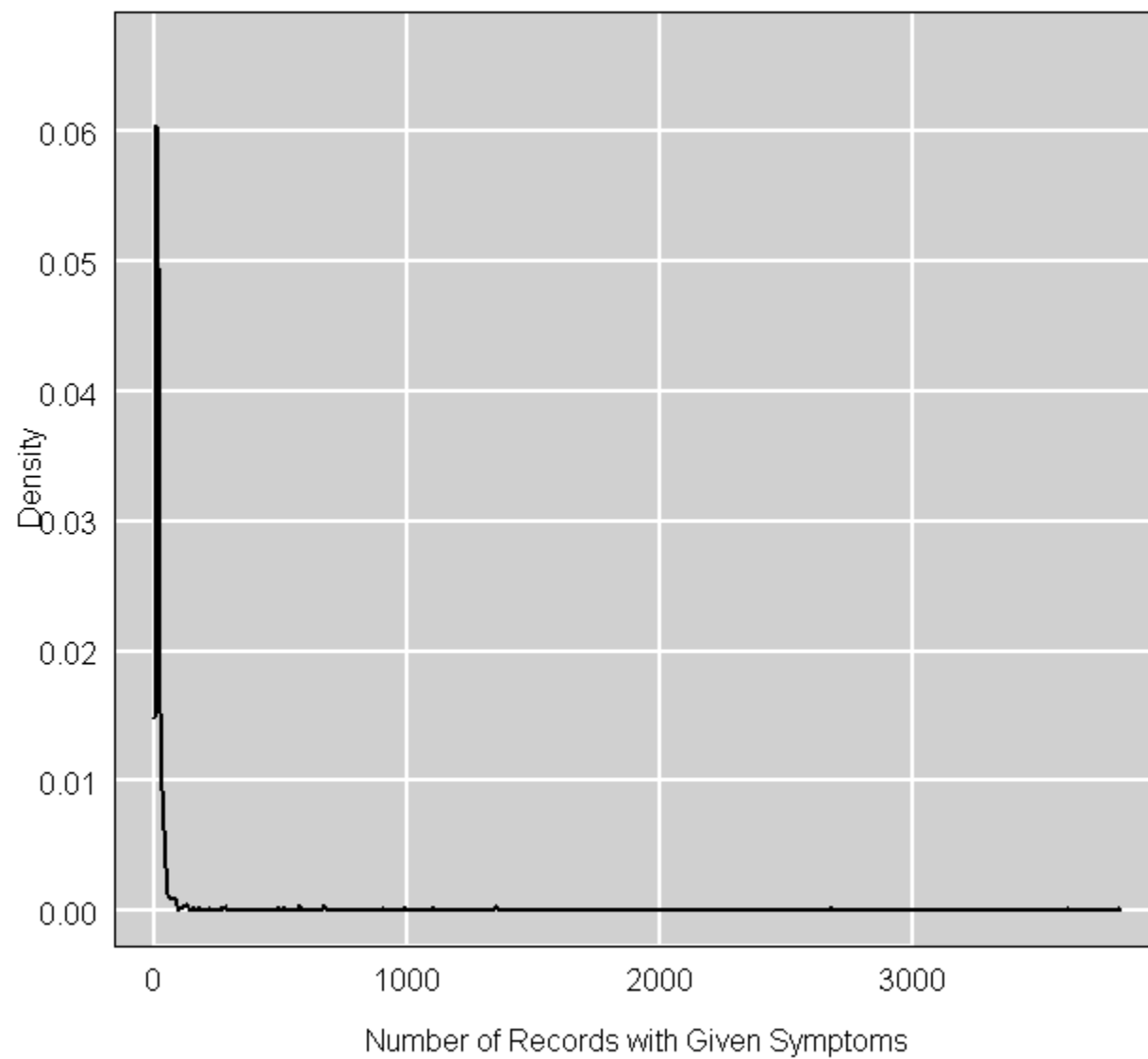
The resulting estimates for drug-symptom combinations are the **number of cases** and the **EBGM**. EBGM stand for empirical Bayes geometric mean. We don't go in the details in this class, but show box plots for the **two variables** for different organ systems. I replace the drug name with "**omitted**".

Bill commented on the inadequacies of attempts to use simplistic data mining association rules for analysis.

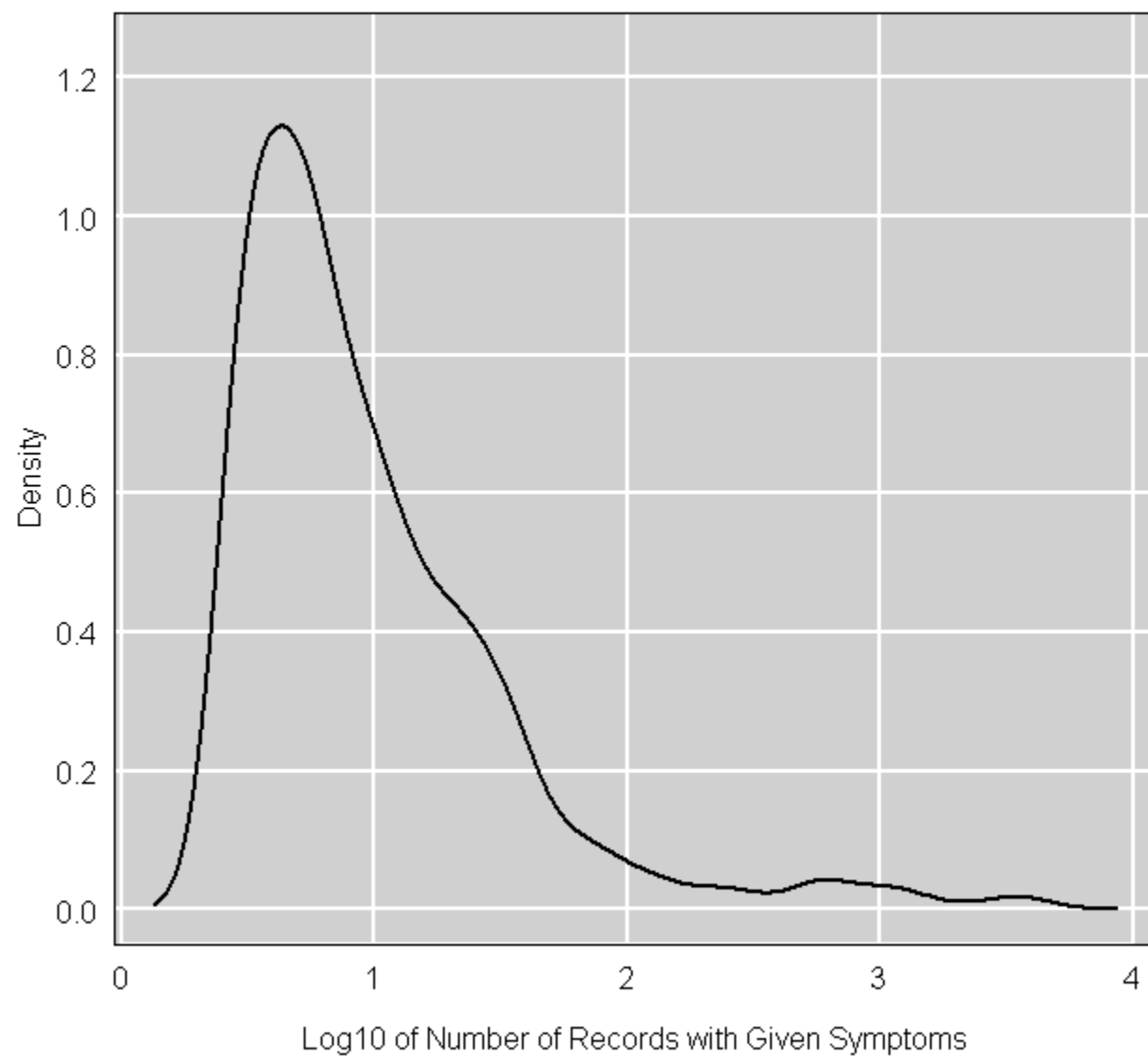
Paper: **Empirical Bayes Screening for Multi-Item Associations**

William DuMouchel and Daryl Pregibon available on the web

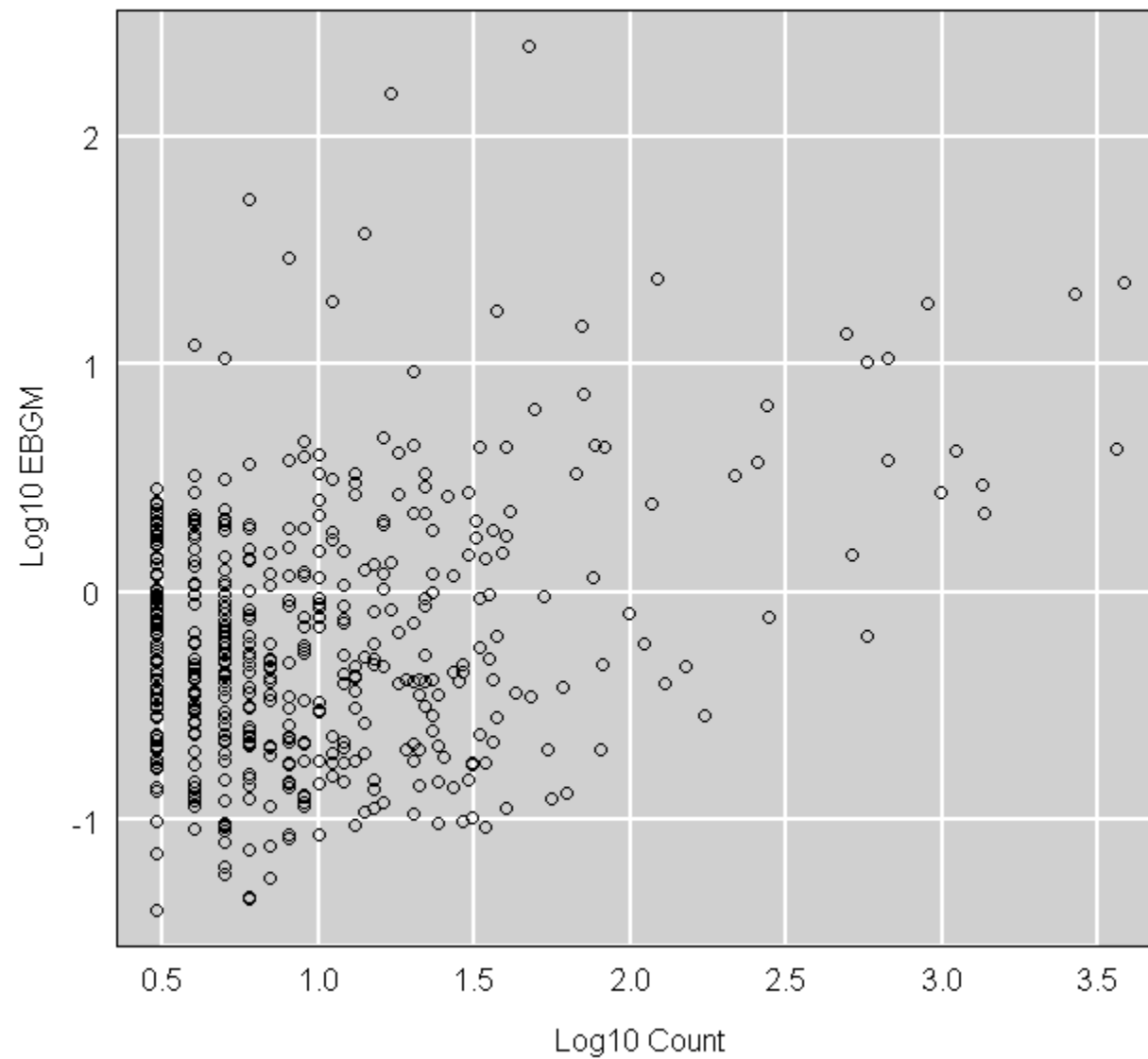
Omitted



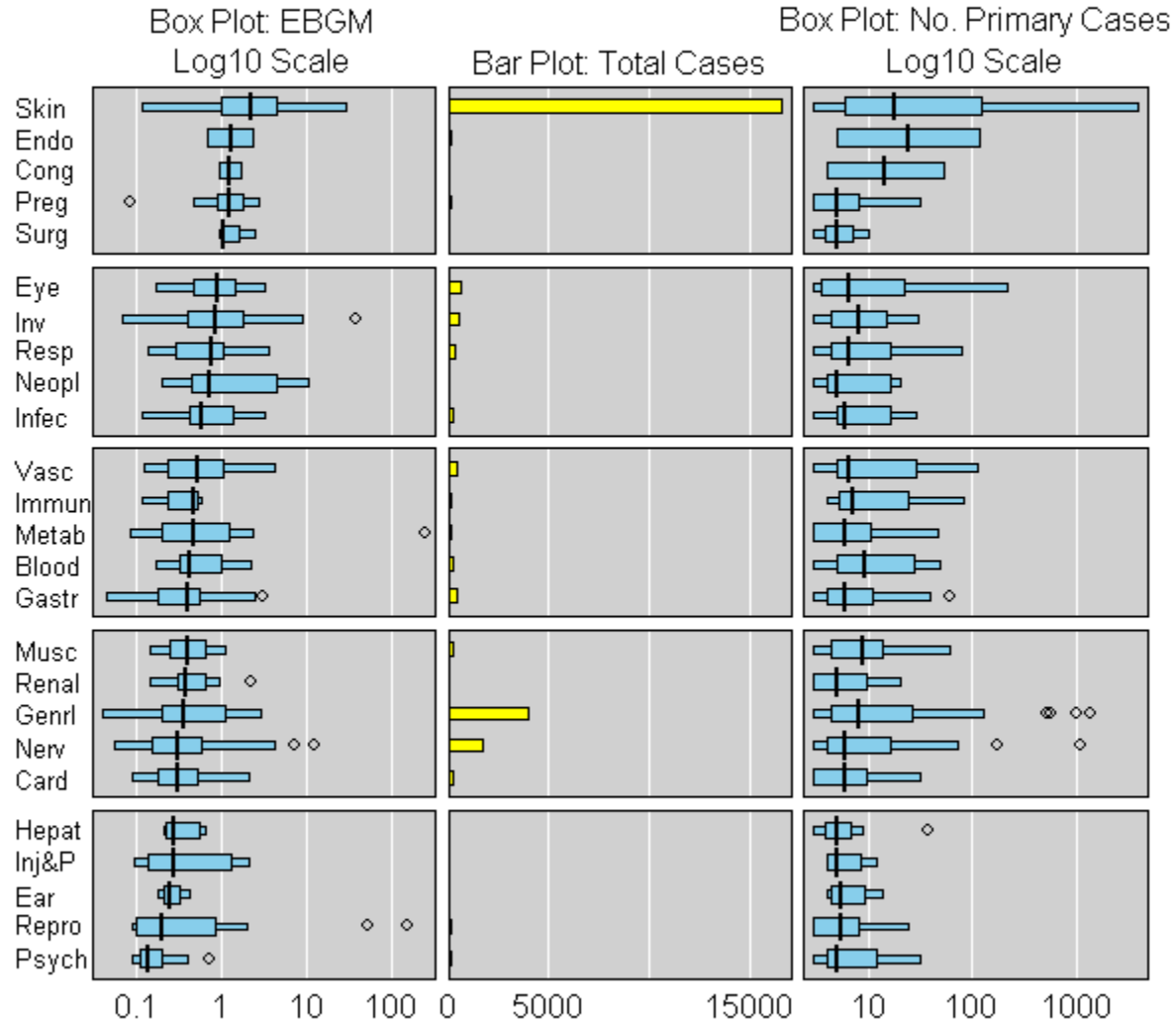
Omitted



Symptom Statistics



Drug Organ System Profile



Box Plots, Large Data Sets and Extensions

Box plots scale pretty well to large data sets

In some cases show more percents would be good

Some distribution can generate very large number of outliers

2D box approximation: My hexagon summary

Start with 2-D hexagon binning of points

Show all occupied hexagon cells: data footprint

Highlight high counts cells containing close to 50% of counts: high density

Apply **gray level erosion** to highlighted cells

Recursively

Removes count from cells based on the **number of exposed cell edges**

Removes **empty cells to expose more edges**

The “median” the last cell standing

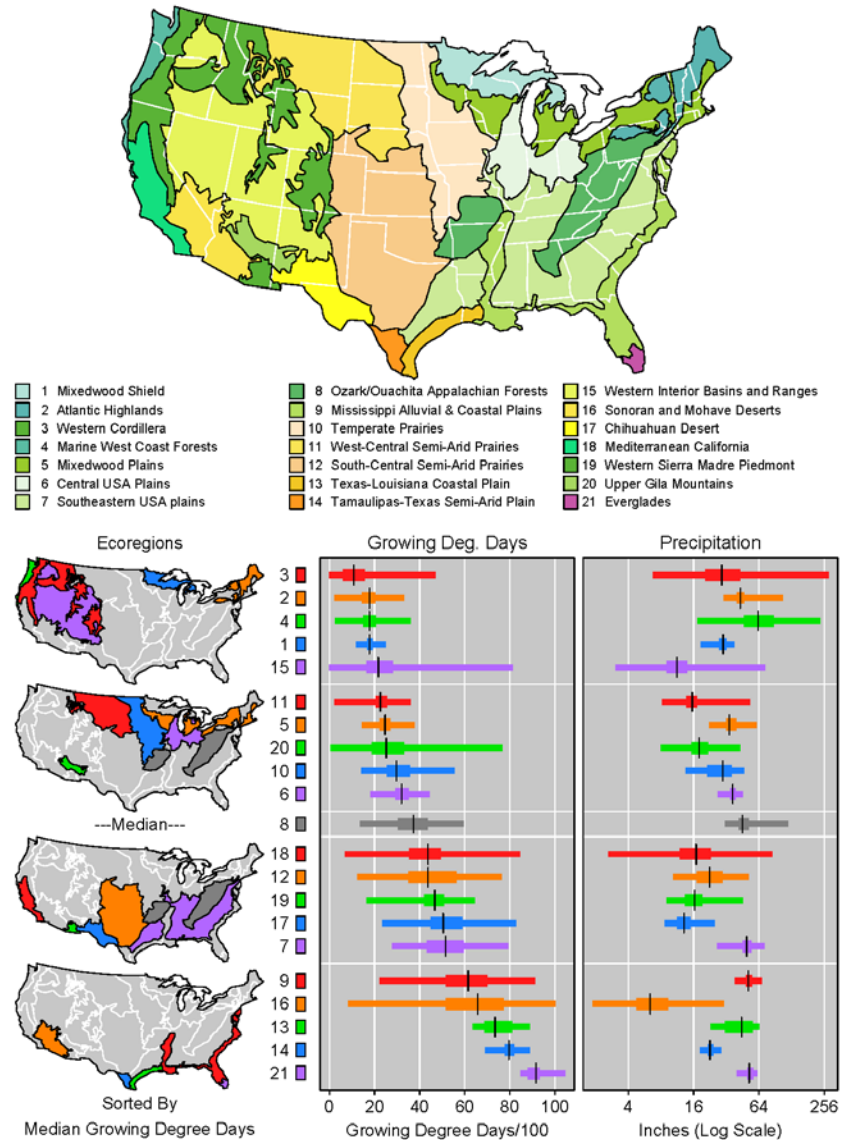
Examples from Carr, Olsen, Pierson, and Courbois [1999]

Data: developed using PRISM (Parameter-elevation Regressions on Independent Slopes Model). PRISM, described by Daly et al (1994).

Roughly 400,000 observations

Bivariate colors: Footprint (gray), highlighted (yellow) median (black)

Figure 1. a=Omernik Ecoregions b=Linked Micromap Boxplots



Numbered Ecoregions Linked by Color to 2-D Binned Summary

Gray: data footprint, Yellow: high density with 50%, Black: median (yellow cell count erosion)

Figure 2: LM Bivariate Boxplots
1961-1990 Precipitation (x) versus Growing Degree Days/100 (y)

