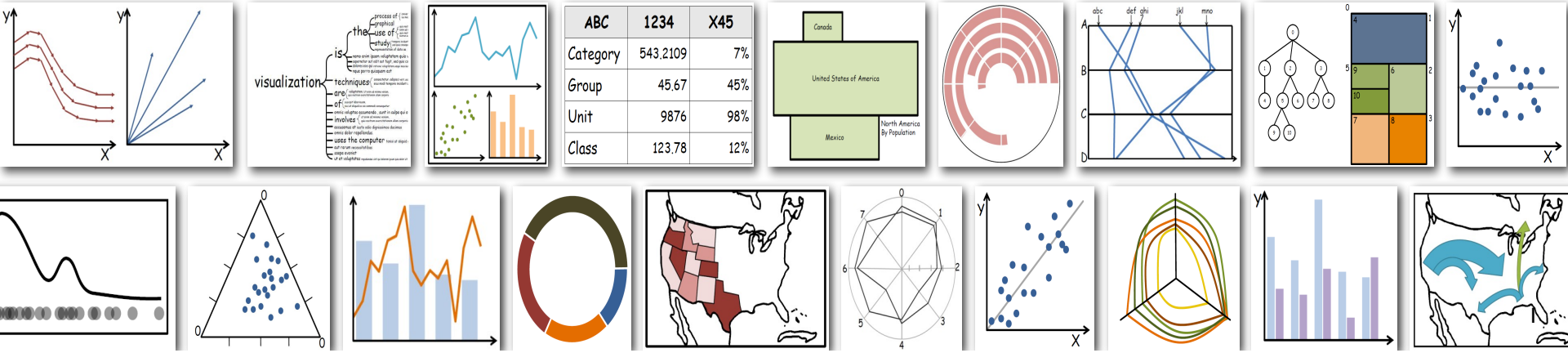


Lecture :

Exploratory Data Analysis

Visualization of 1D/2D Data

DATA ANALYSIS & VISUALIZATION

FALL 2021

Dr. Muhammad Faisal Cheema

FAST NU

Outline

- EDA
- Visualization / Analysis
  - One variable
  - Two variables
  - More than two variables
  - Other types of data
  - Dimension reduction

EDA and Visualization

- Exploratory Data Analysis (EDA) and Visualization are very important steps in any analysis task.

- get to know your data!
  - distributions (symmetric, normal, skewed)
  - data quality problems
  - outliers
  - correlations and inter-relationships
  - subsets of interest
  - suggest functional relationships

- Sometimes EDA or viz might be the goal!

Exploratory Data Analysis (EDA)

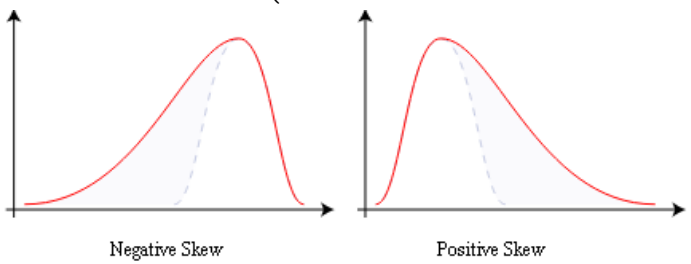
- Goal: get a general sense of the data
  - means, medians, quantiles, histograms, boxplots
  - You should always look at every variable - you will learn something!
- data-driven (model-free)
- Think interactive and visual
  - Humans are the best pattern recognizers
  - You can use more than 2 dimensions!
    - x,y,z, space, color, time….

- Especially useful in early stages of data mining
  - detect outliers (e.g. assess data quality)
  - test assumptions (e.g. normal distributions or skewed?)
  - identify useful raw data & transforms (e.g. log(x))

- Bottom line: it is always well worth looking at your data!

Summary Statistics

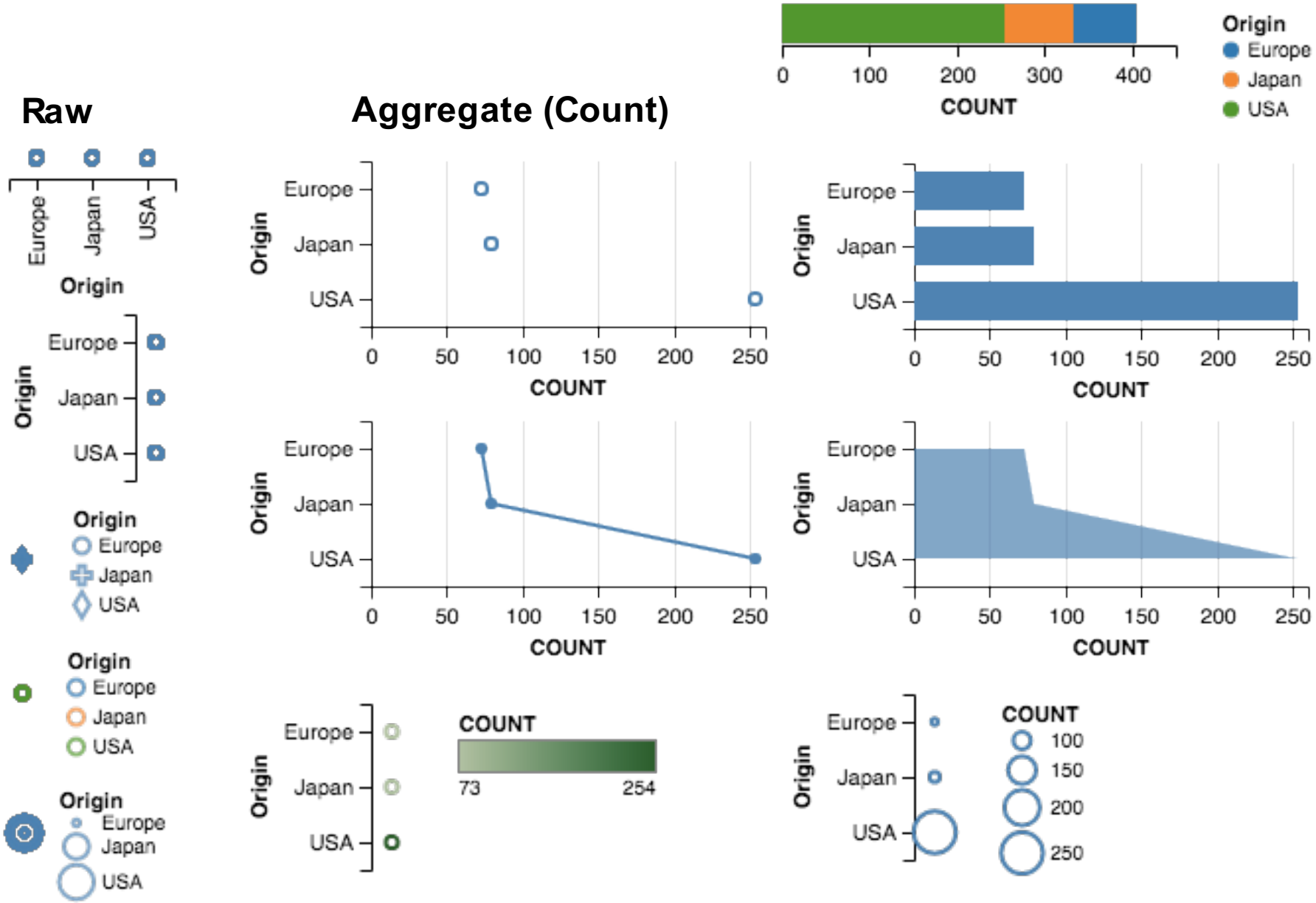
- *not* visual
- sample statistics of data X
  - mean: $\mu = \sum_i X_i / n$
  - mode: most common value in X
  - median: $\mathbf{X}$=sort(X), median = $\mathbf{X}_{n/2}$ (half below, half above)
  - quartiles of sorted $\mathbf{X}$: Q1 value = $\mathbf{X}_{0.25n}$, Q3 value = $\mathbf{X}_{0.75\ n}$
    - interquartile range: value(Q3) - value(Q1)
    - range: max(X) - min(X) = $\mathbf{X}_n - \mathbf{X}_1$
  - variance: $\sigma^2 = \sum_i (X_i - \mu)^2 / n$
  - skewness: $\sum_i (X_i - \mu)^3 \ / \ [(\sum_i (X_i - \mu)^2)^{3/2}]$
    - zero if symmetric; right-skewed more common (what kind of data is right skewed?)



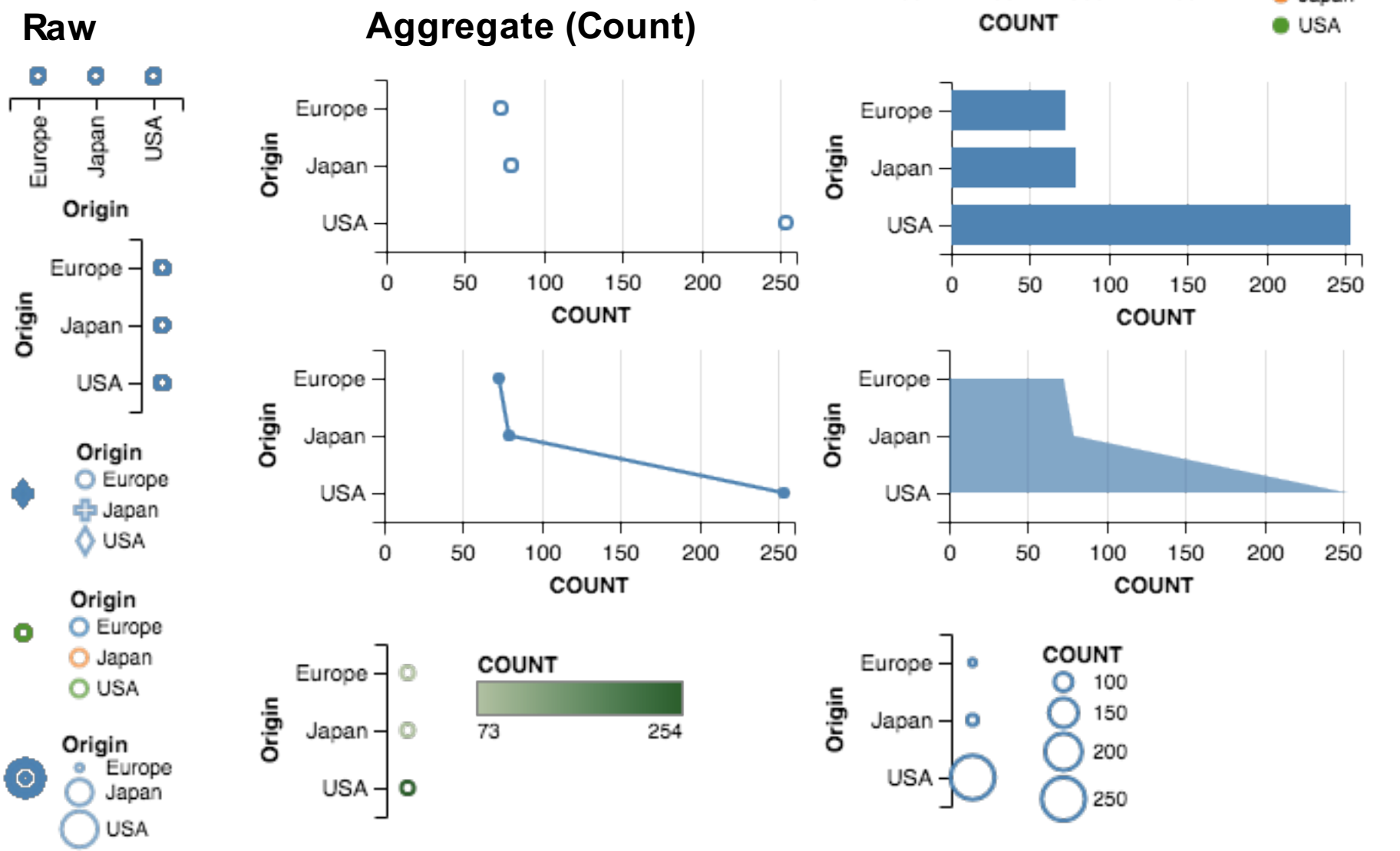
  - number of distinct values for a variable (see unique() in R)
  - Don't need to report all of thses: Bottom line…do these numbers make sense???

Univariate (1D) Analysis / Single Variable Visualizations

1D: Nominal

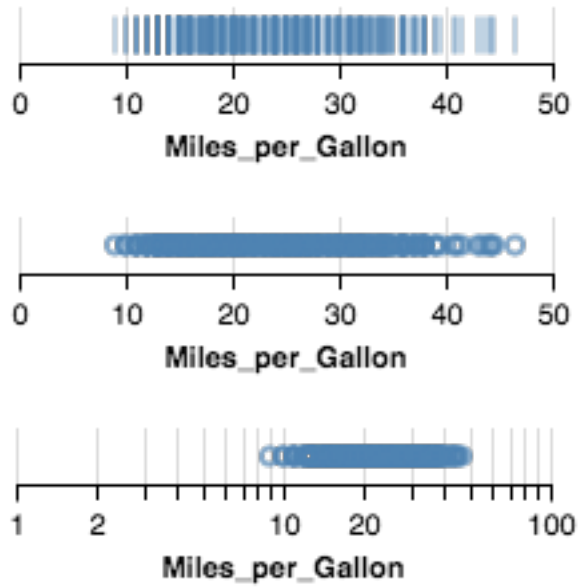

Raw
Aggregate (Count)
Origin
Europe
Japan
USA
COUNT
0
100
200
300
400
50
150
250
73
254

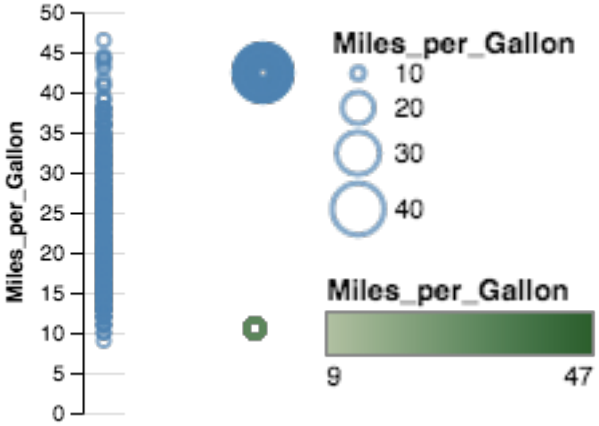
Expressive? Effective?



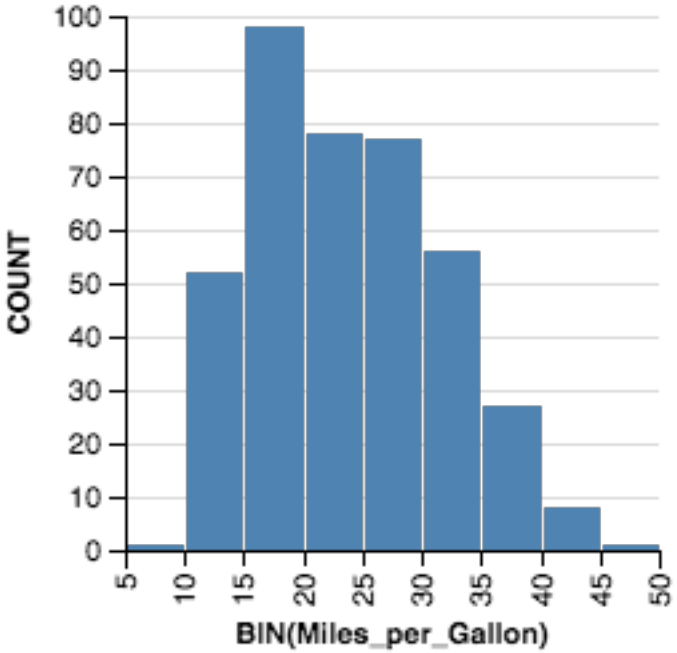
1D: Quantitative

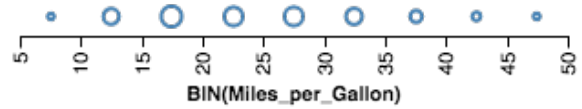
Raw

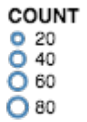


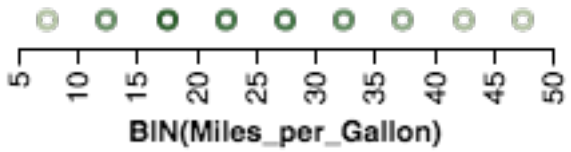


Aggregate (Count)



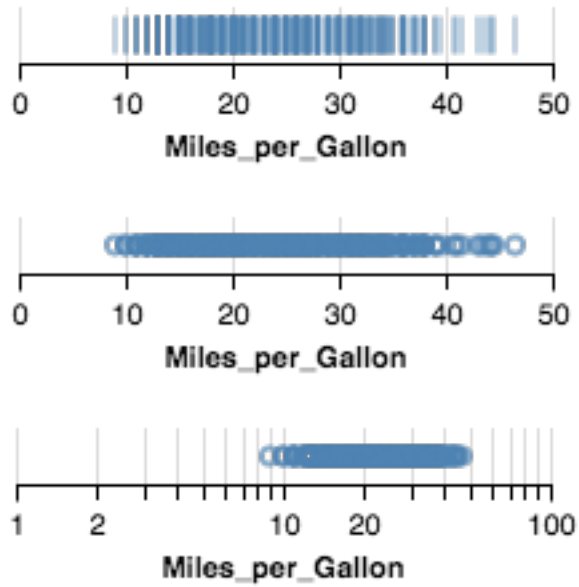


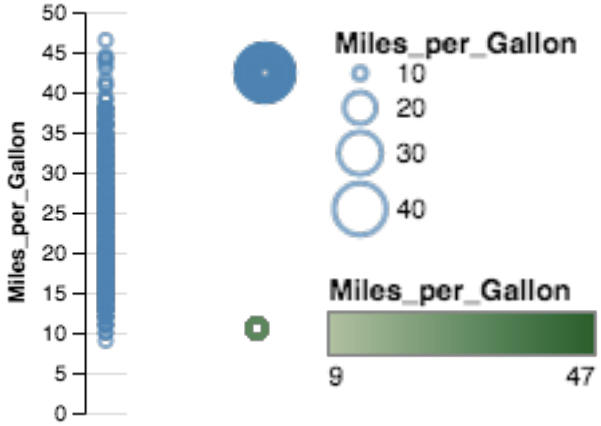




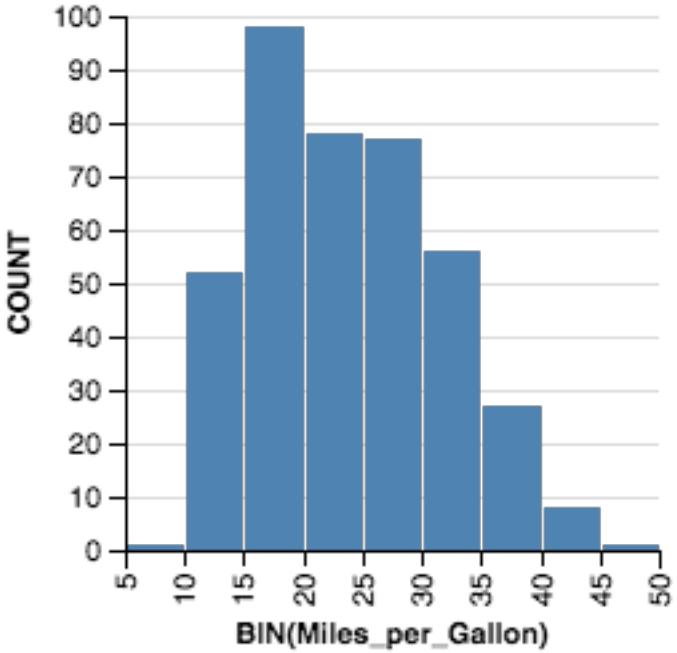
Expressive? Effective?

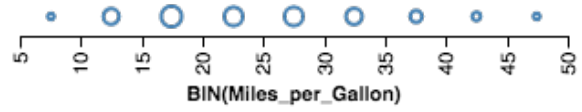
Raw





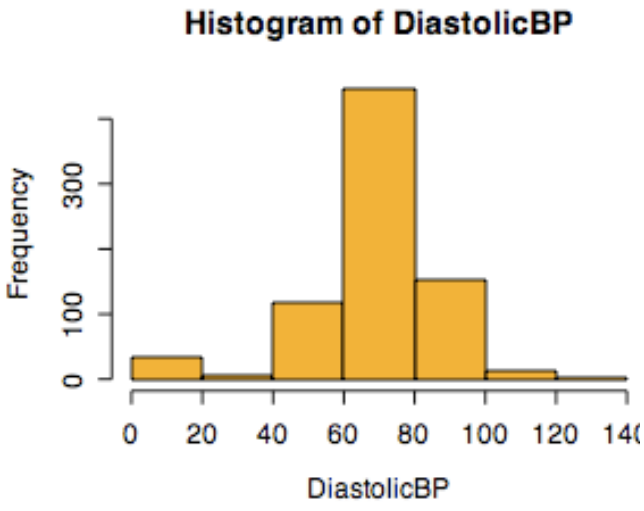
Aggregate (Count)

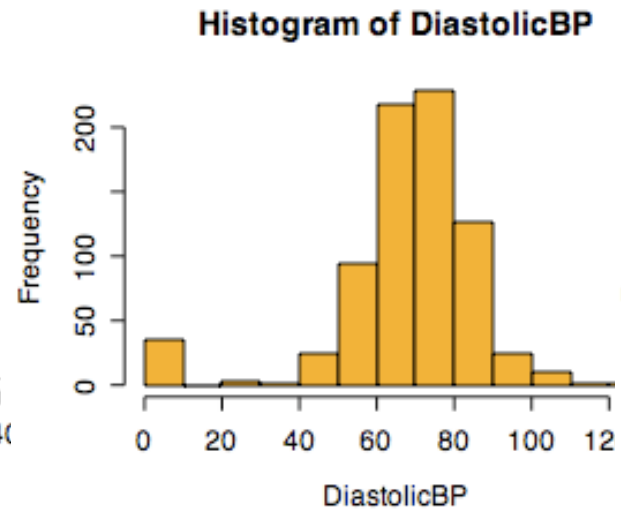


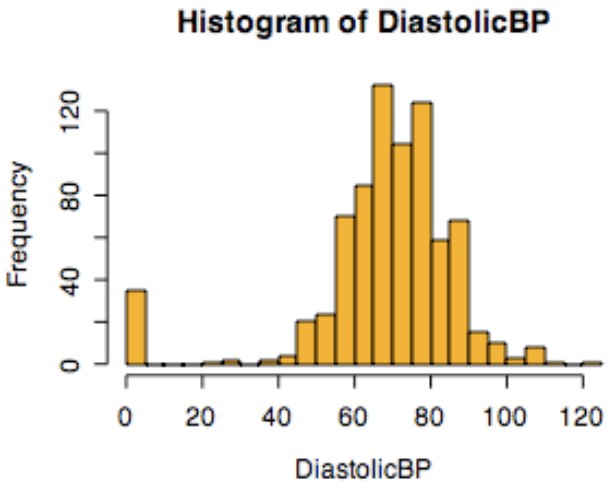


Histograms – A popular viz for 1D data analysis

- Shows center, variability, skewness, modality,
- outliers, or strange patterns.
- Bin width and position matter
- Beware of real zeros



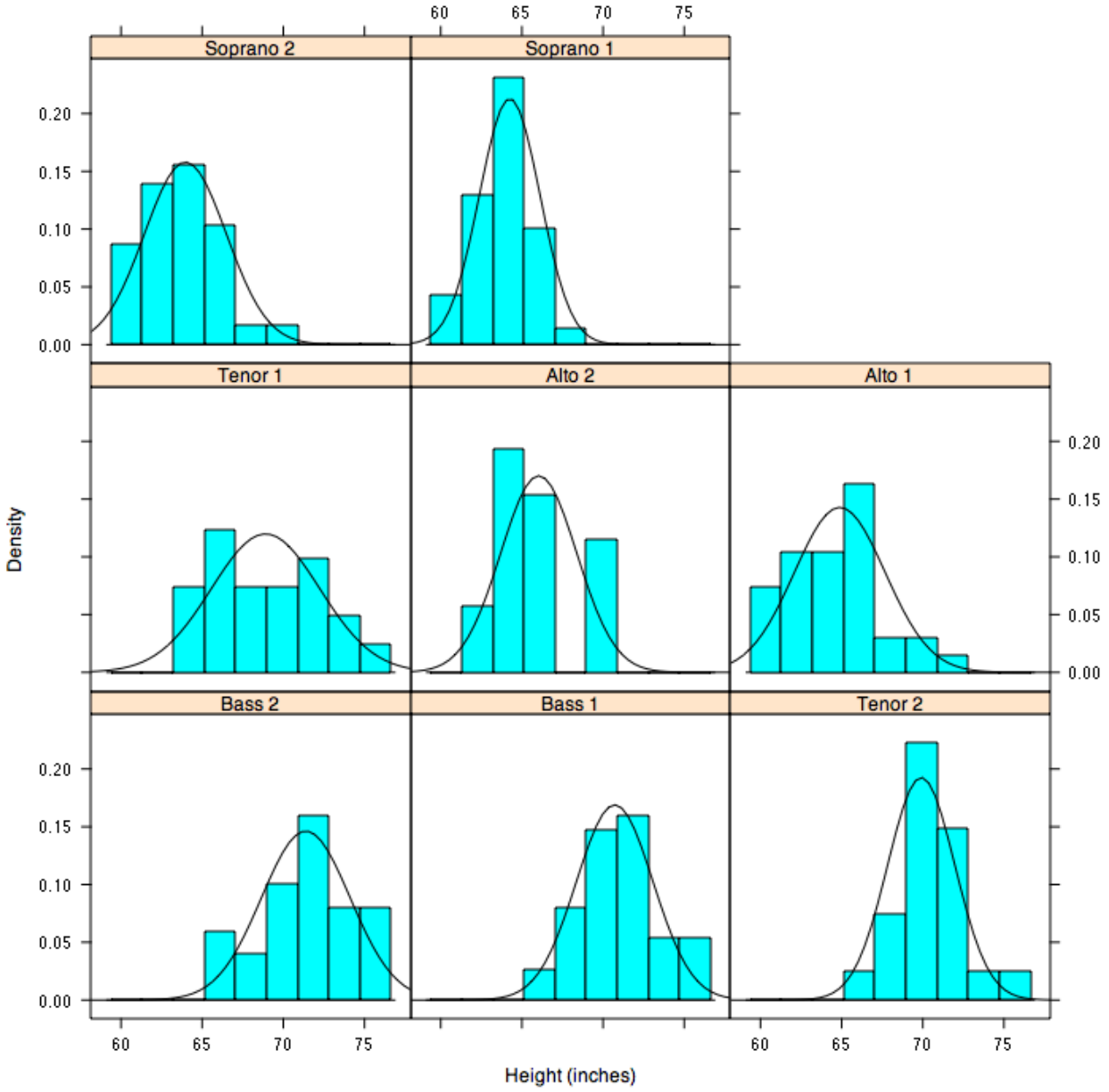




Issues with Histograms

- For small data sets, histograms can be misleading.
  - Small changes in the data, bins, or anchor can deceive
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.
- Histograms effectively only work with 1 variable at a time
  - But 'small multiples' can be effective

But be careful with axes and scales!

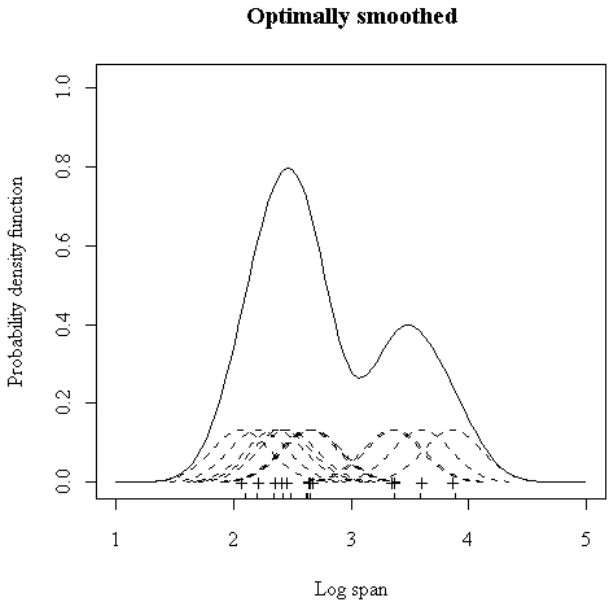


Smoothed Histograms - Density Estimates

- Kernel estimates smooth out the contribution of each datapoint over a local neighborhood of that point.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

h is the kernel width

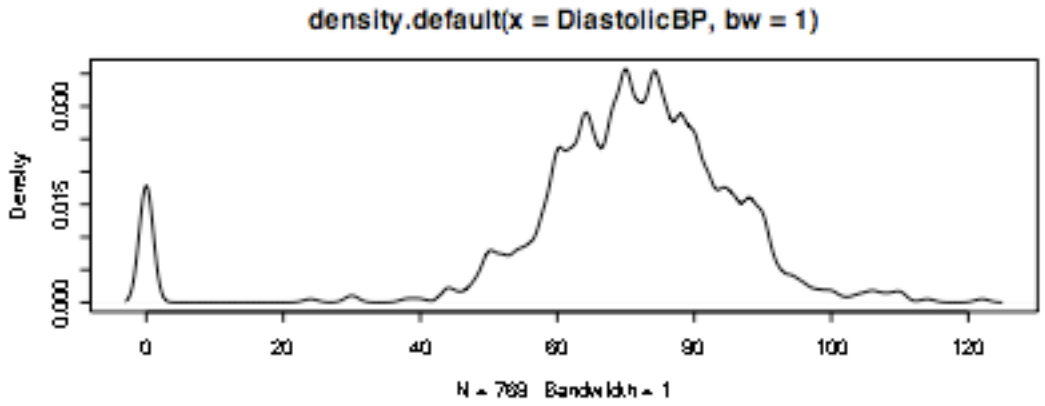


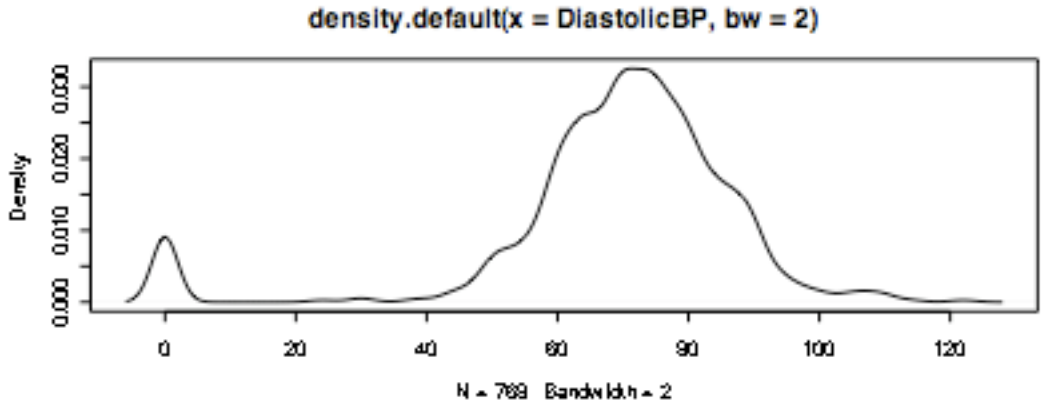
- Gaussian kernel is common:

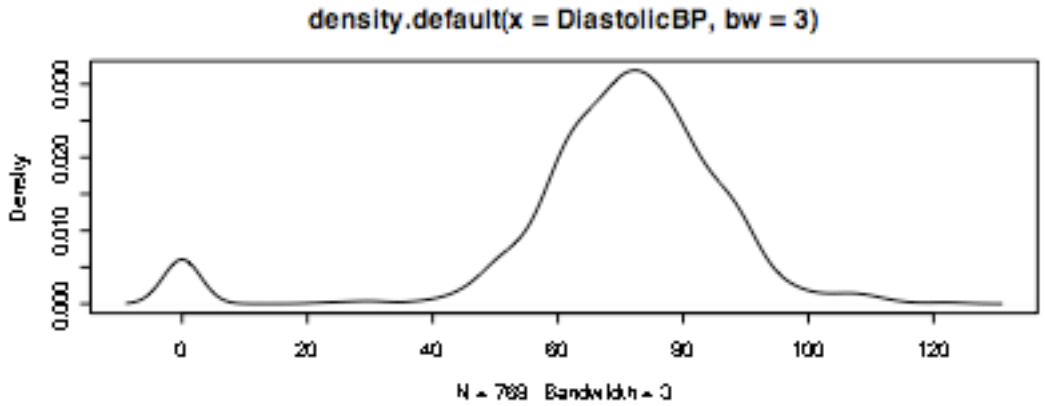
$$Ce^{-\frac{1}{2}\left(\frac{x-x(i)}{h}\right)^2}$$

Bandwidth choice is an art

Usually want to try several

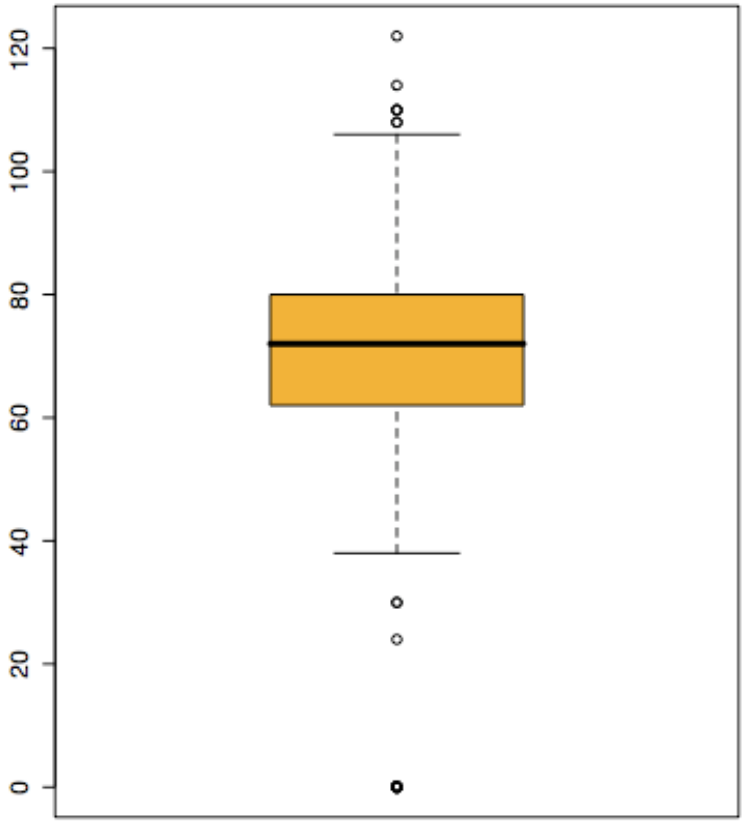






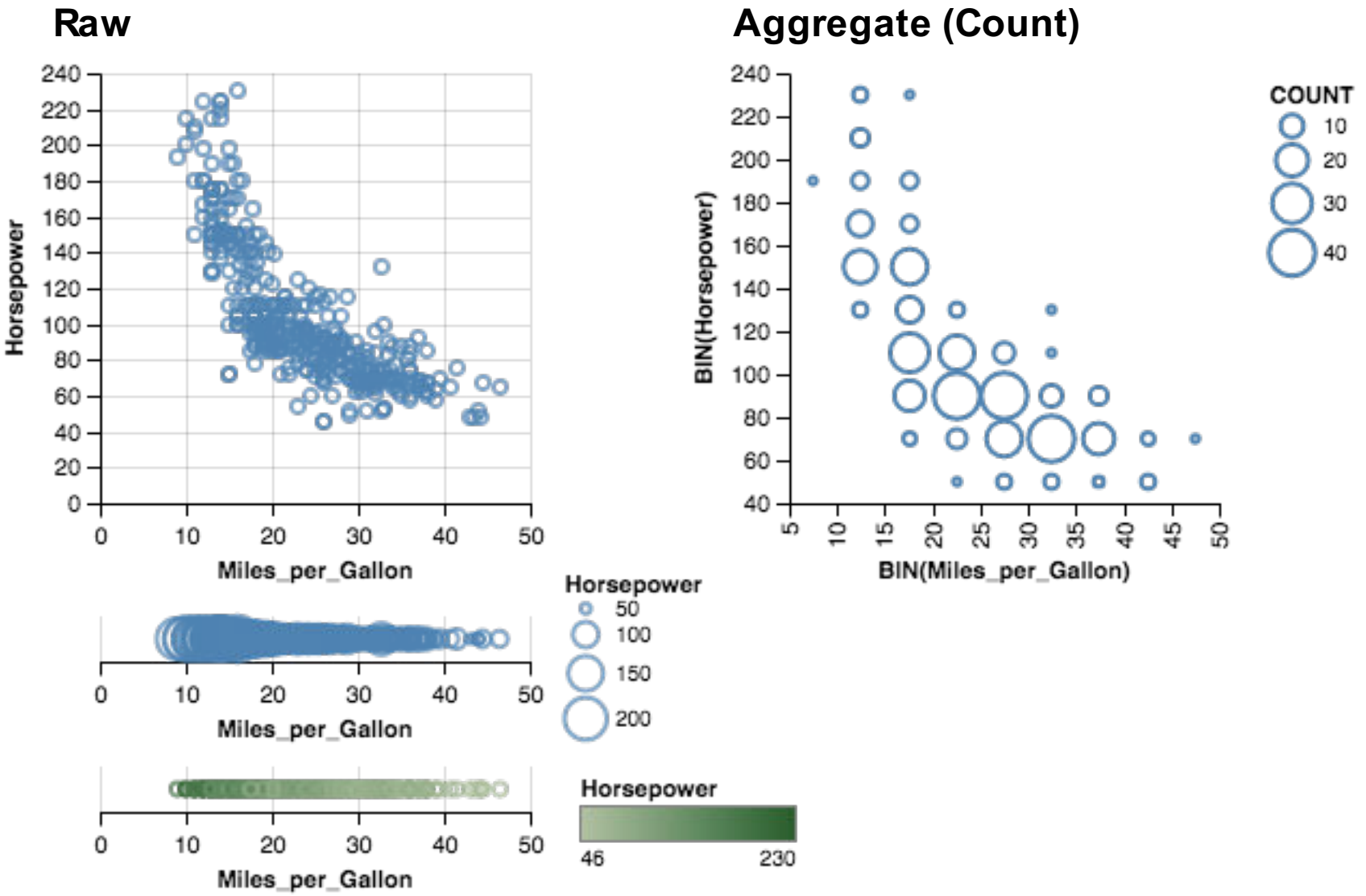
Boxplots – Another popular viz for 1D data analysis

- Shows a lot of information about a variable in one plot
  - Median
  - IQR
  - Outliers
  - Range
  - Skewness
- Negatives
  - Overplotting
  - Hard to tell distributional shape
  - no standard implementation in software (many options for whiskers, outliers)

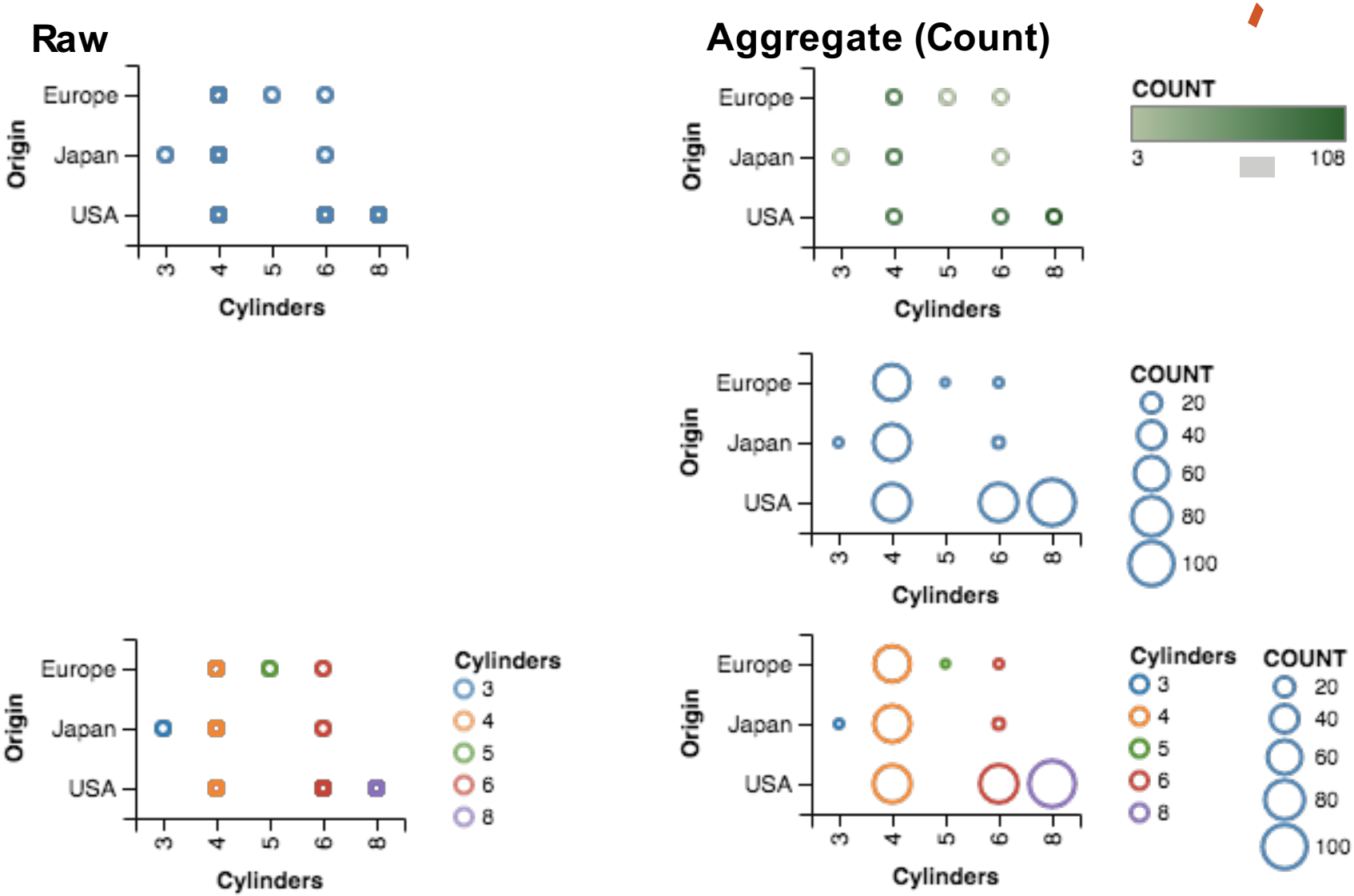


Bivariate (2D) Analysis / Visualizations

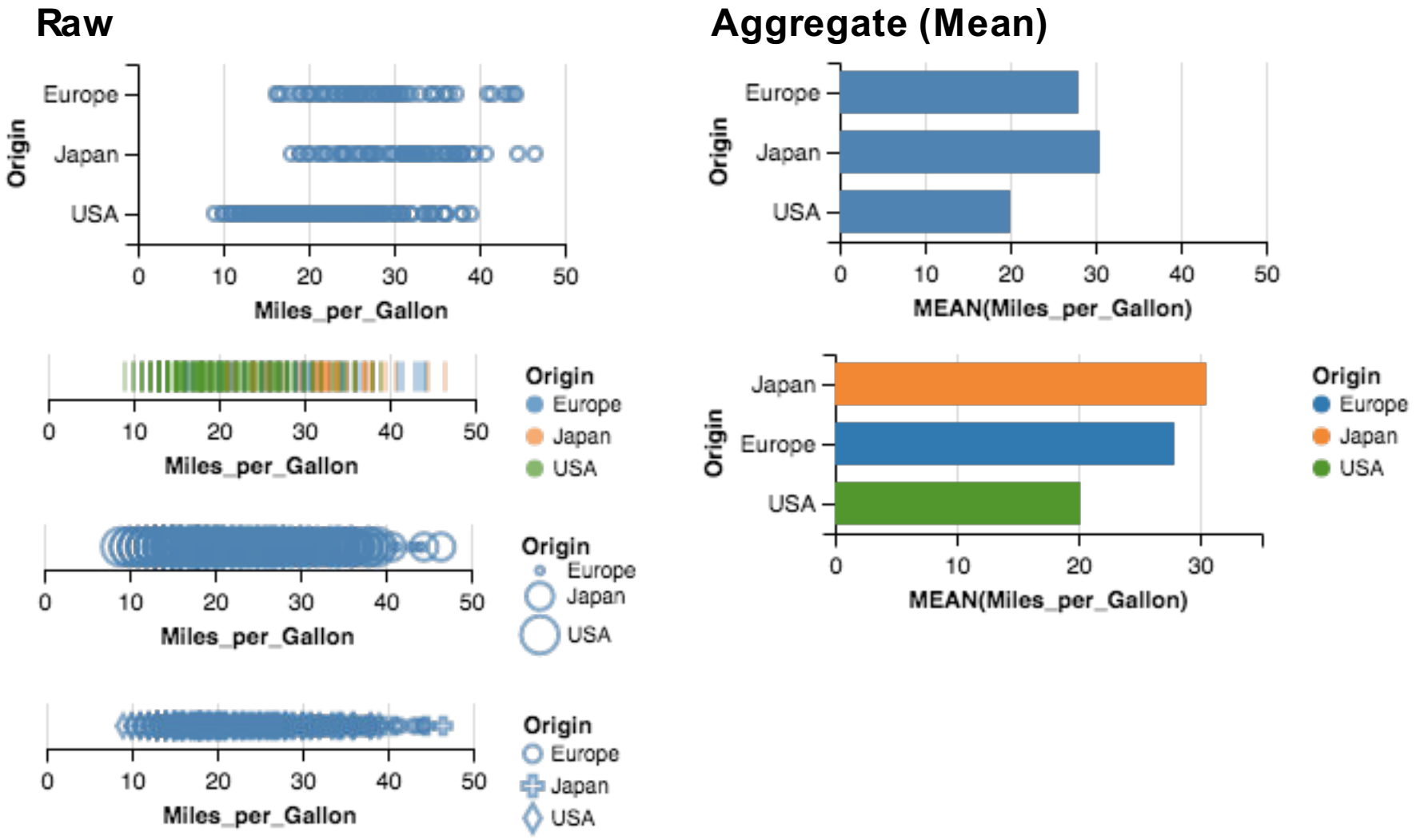
2D: Quantitative x Quantitative



2D: Nominal x Nominal

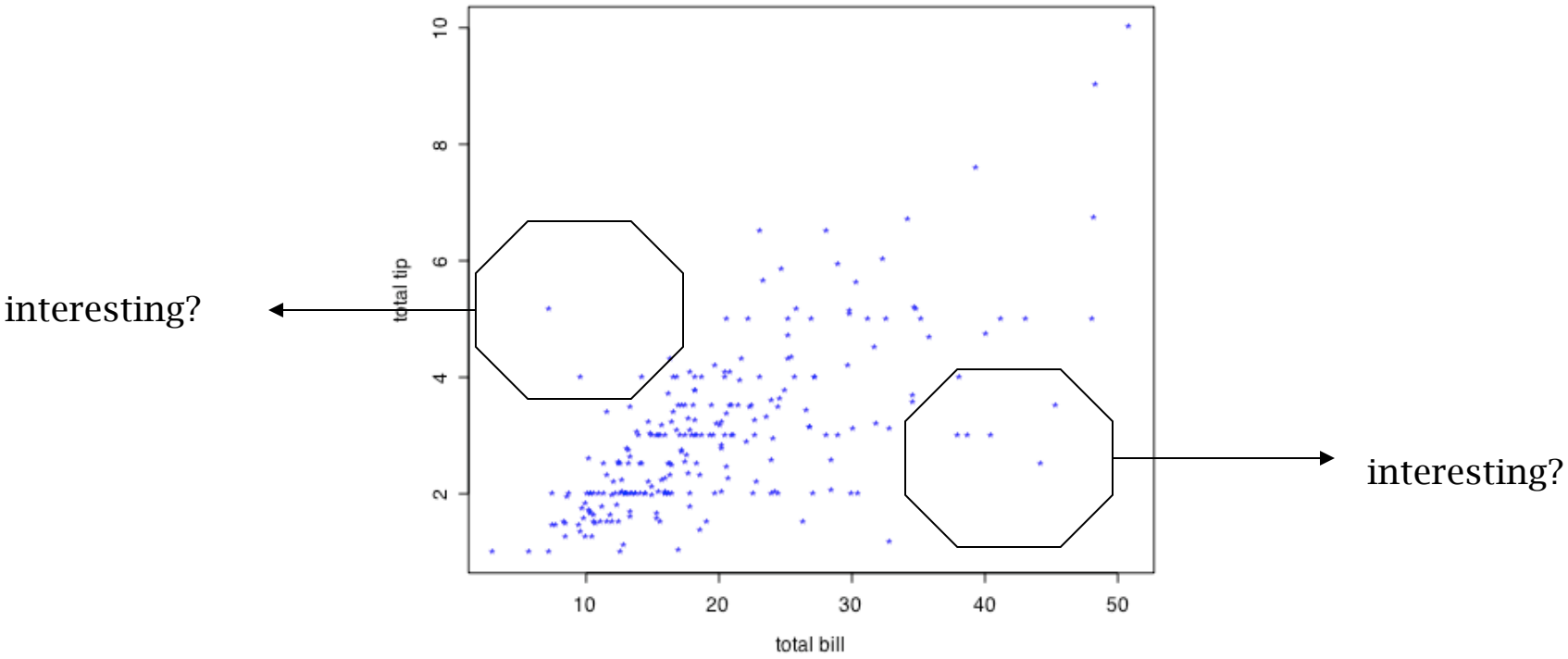


2D: Nominal x Quantitative



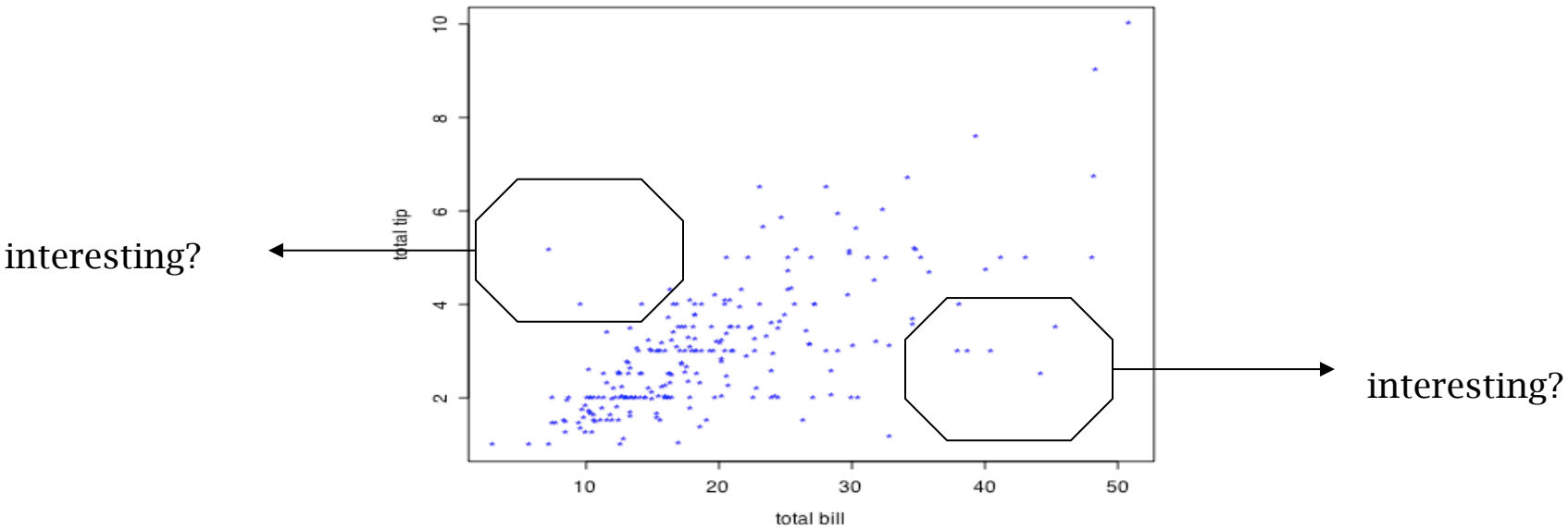
Two Continuous Variables

- For two numeric variables, the scatterplot is the obvious choice

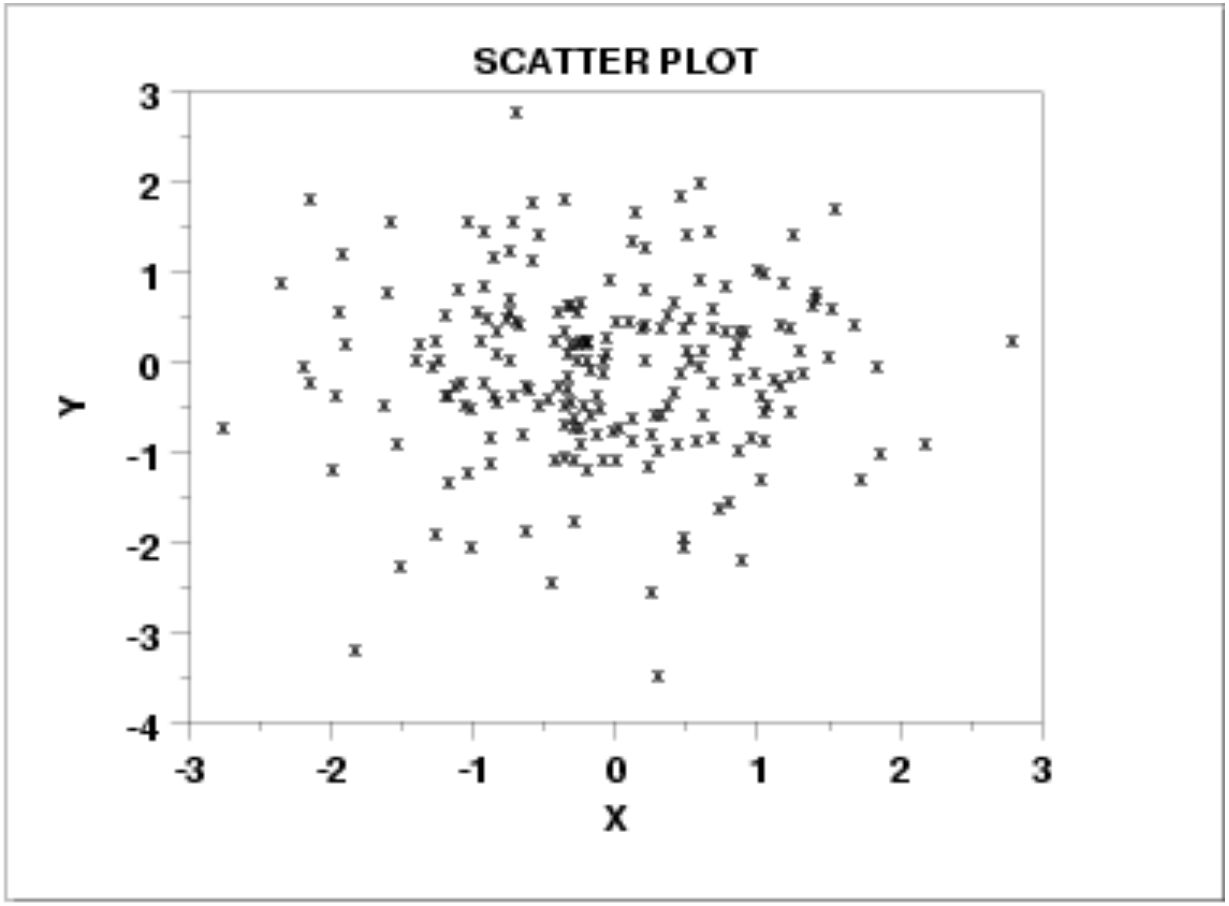


2D Scatterplots

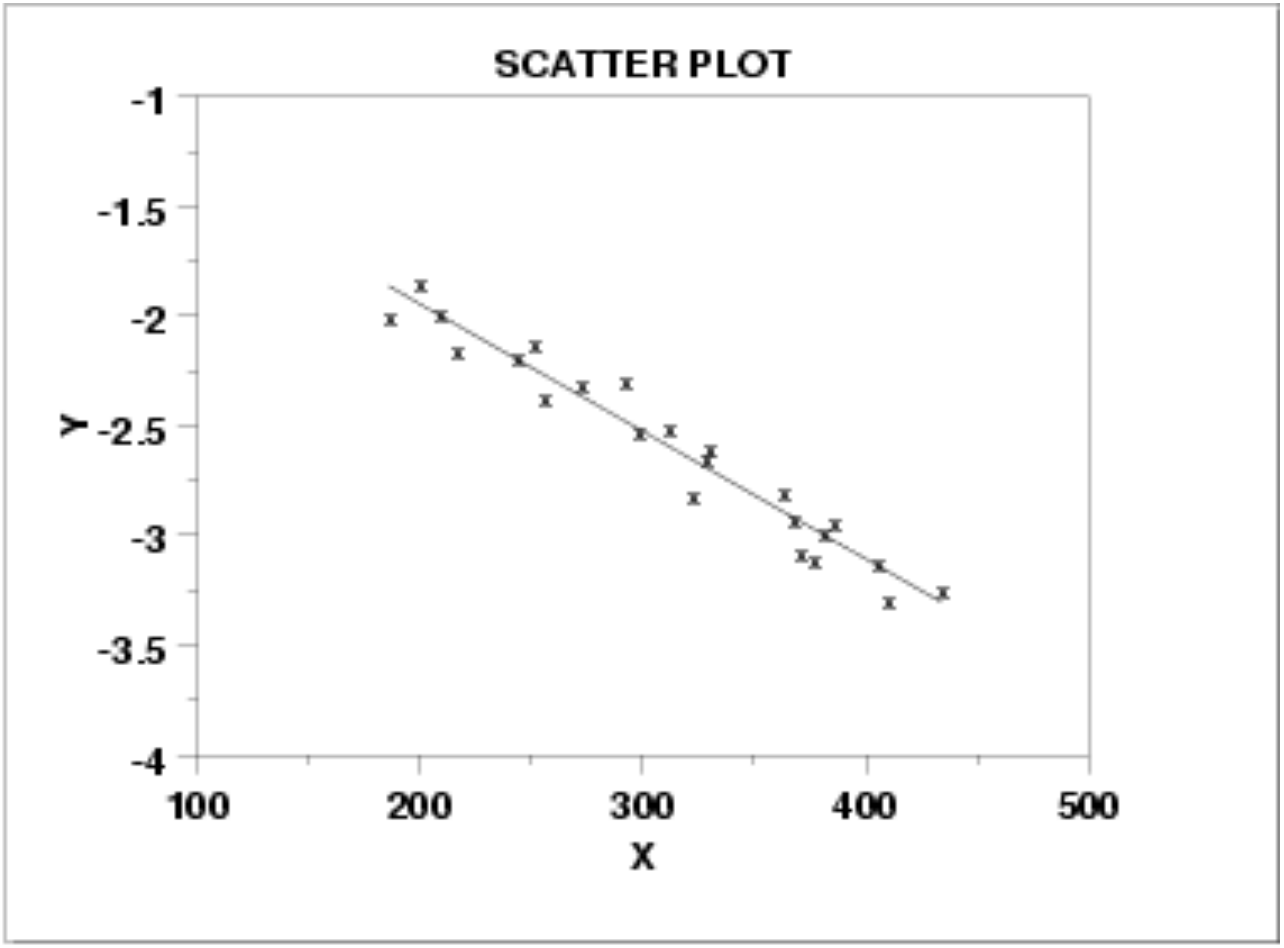
- standard tool to display relation between 2 variables
  - e.g. y-axis = response, x-axis = suspected indicator
- useful to answer:
  - x,y related?
    - linear
    - quadratic
    - other
  - variance(y) depend on x?
  - outliers present?



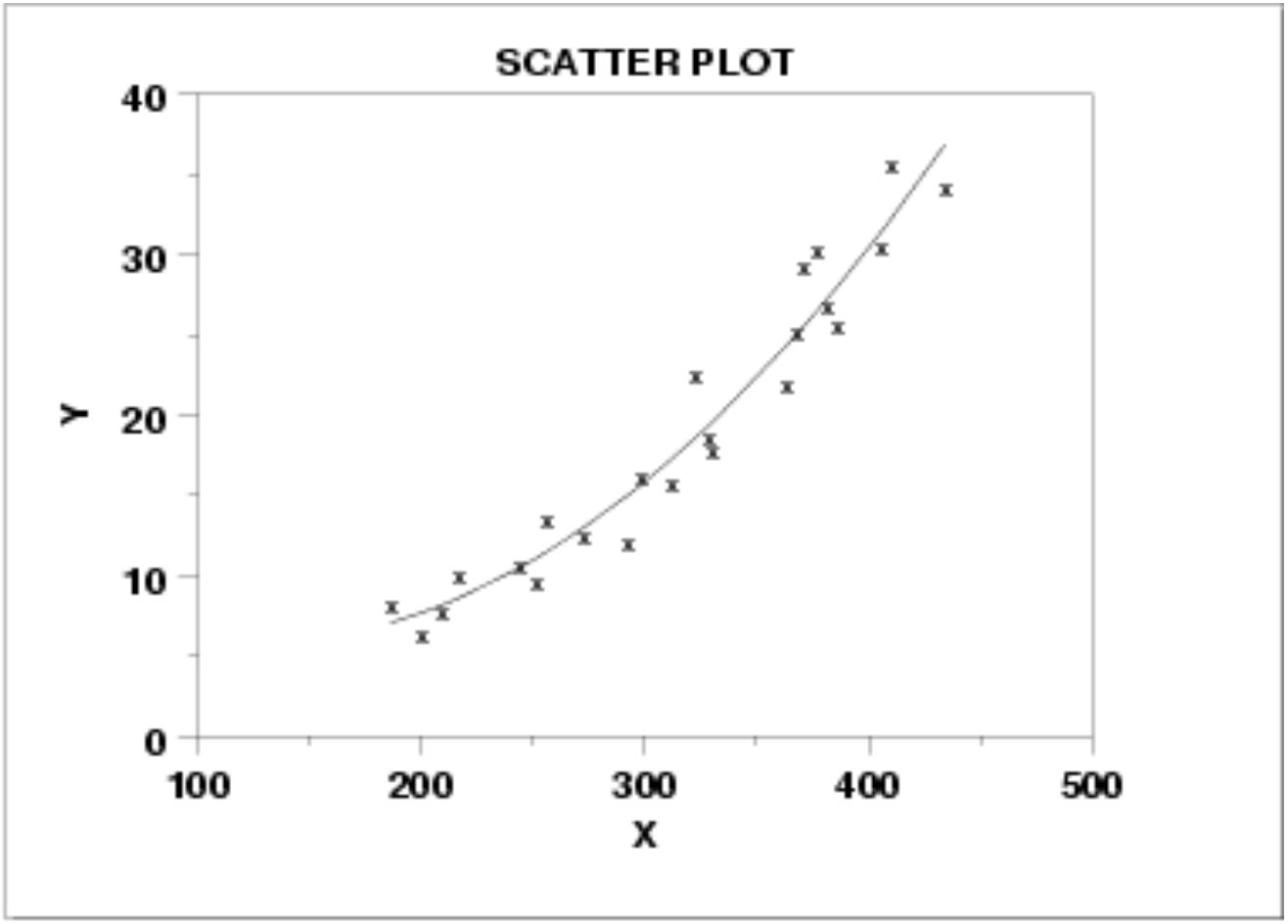
Scatter Plot: No apparent relationship



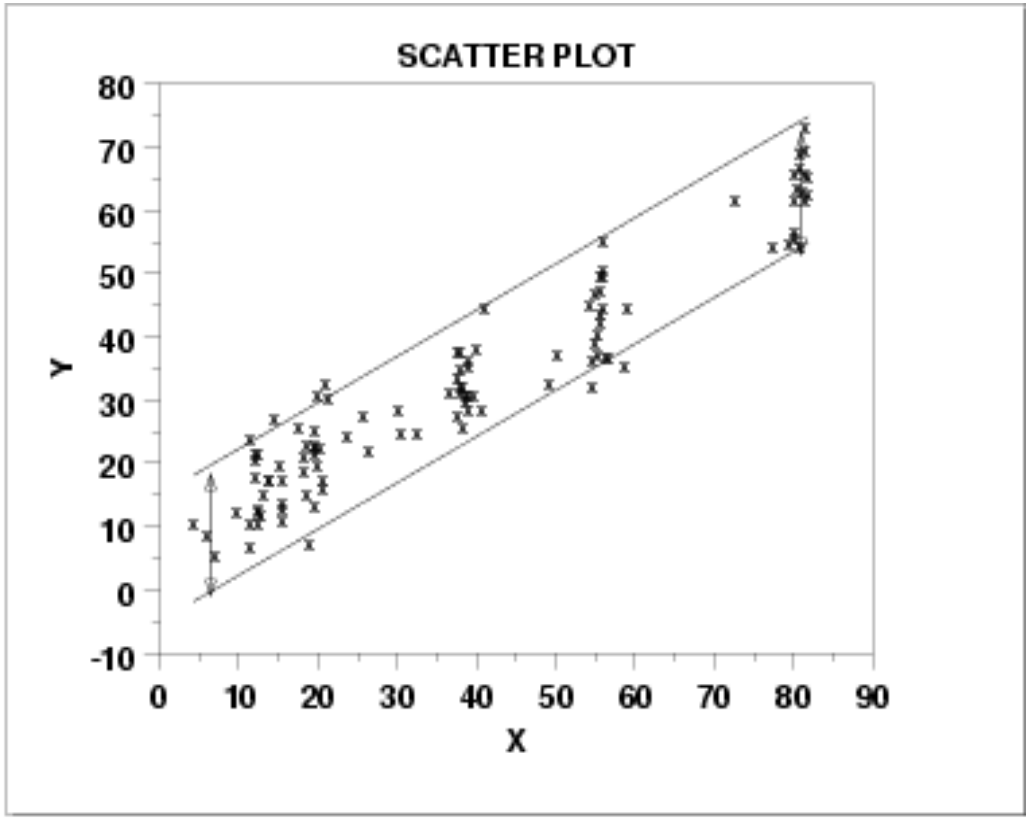
Scatter Plot: Linear relationship



Scatter Plot: Quadratic relationship

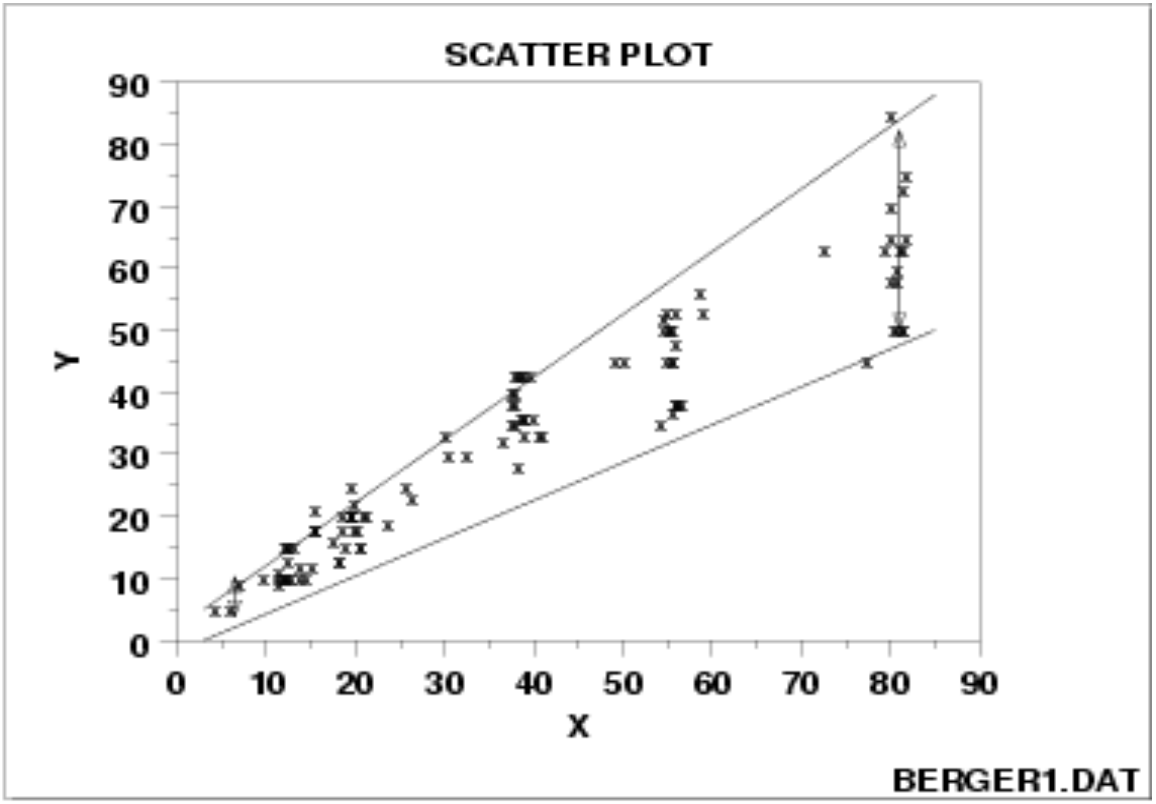


Scatter plot: Homoscedastic



Why is this important in classical statistical modelling?

Scatter plot: Heteroscedastic



variation in *Y* differs depending on the value of *X*
e.g., Y = annual tax paid, X = income

Two variables - continuous

- Scatterplots
  - But can be bad with lots of data

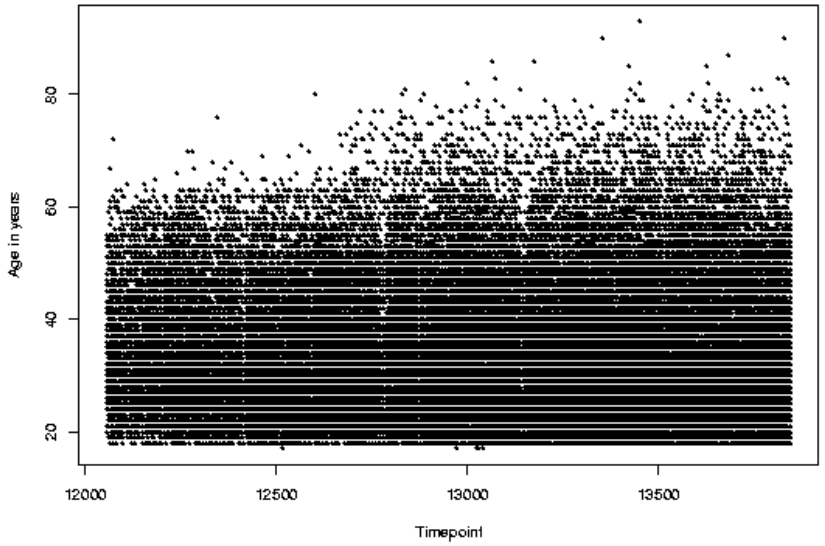
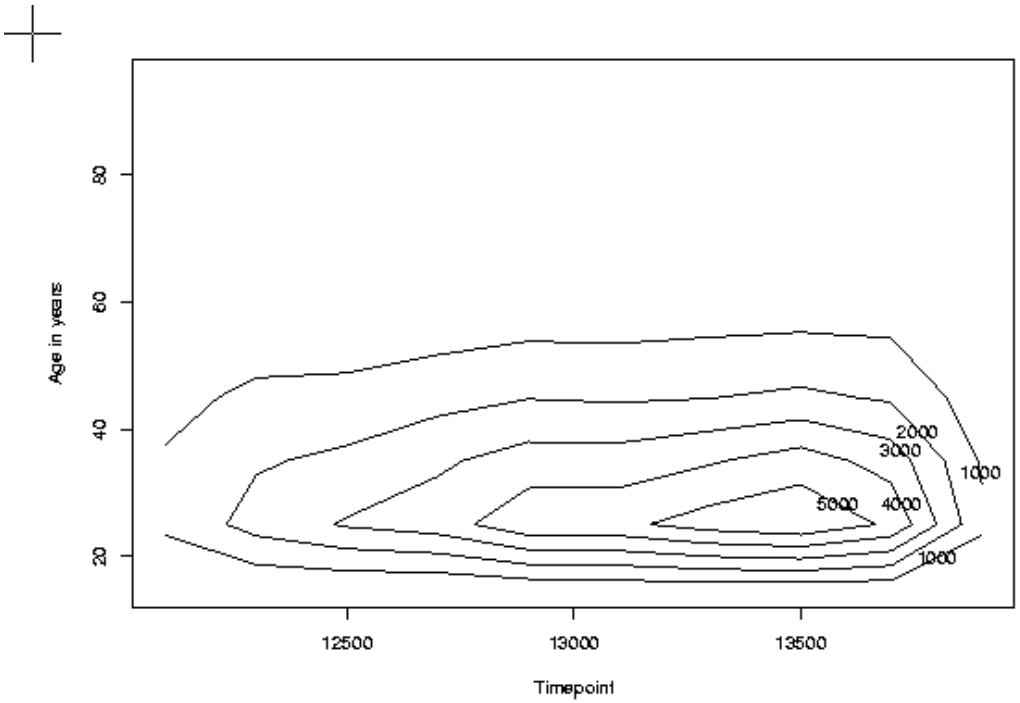


Figure 3.7: A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.

Two variables - continuous

- What to do for large data sets
  - Contour plots

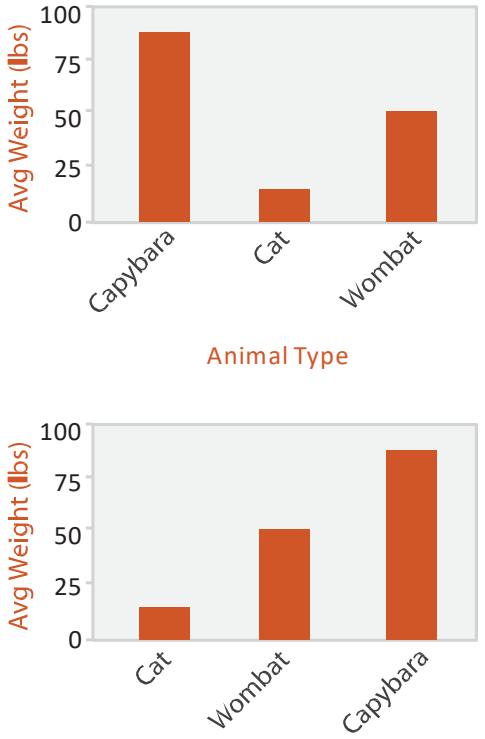


Separate, Order, and Align: Categorical Regions

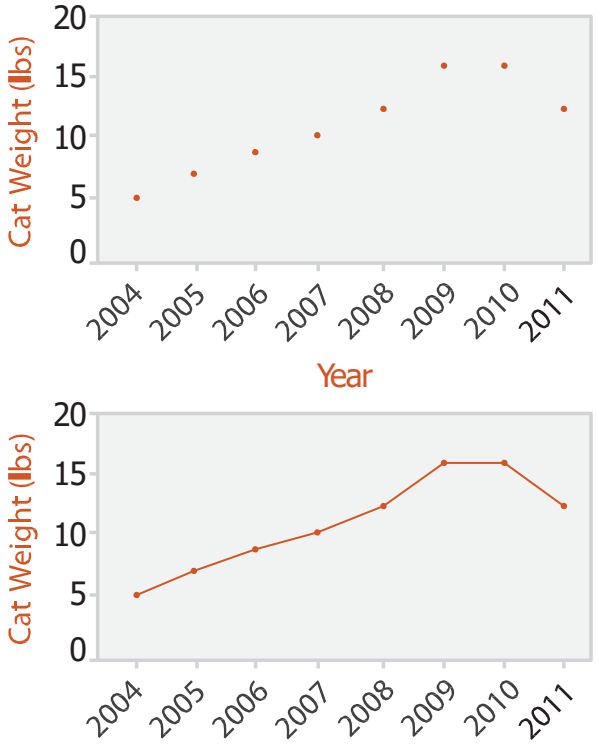
- Categorical: =, !=
- Spatial position can be used for categorical attributes
- Use regions, distinct contiguous bounded areas, to encode categorical attributes
- Three operations on the regions:
  - Separate (use categorical attribute)
  - Align (use some other ordered attribute)
  - Order
- Alignment and order can use same or different attribute

List Alignment: Bar Charts

- Data: one quantitative attribute, one categorical attribute
- Task: lookup & compare values
- How: line marks, vertical position (quantitative), horizontal position (categorical)
- What about length?
- Ordering criteria: alphabetical or using quantitative attribute
- Scalability: distinguishability
  - bars at least one pixel wide
  - hundreds



Dot and Line Charts

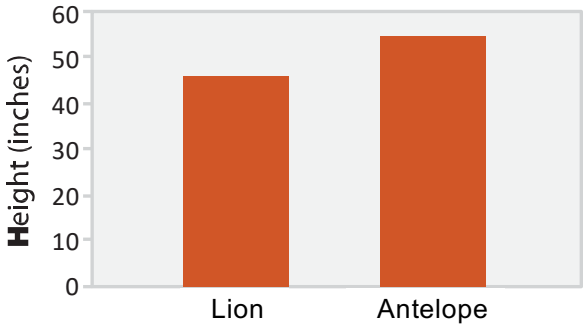


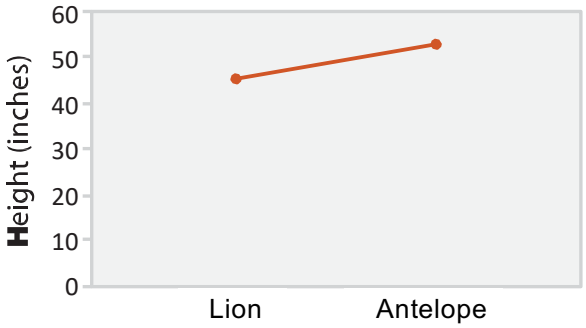
- Data: one quantitative attribute, one ordered attribute
- Task: lookup values, find outliers and trends
- How: point mark and positions

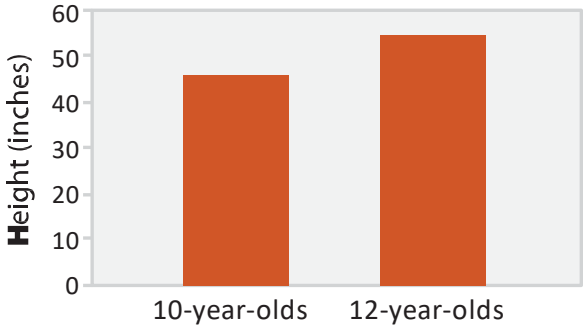
- Line Charts: add connection mark (line)

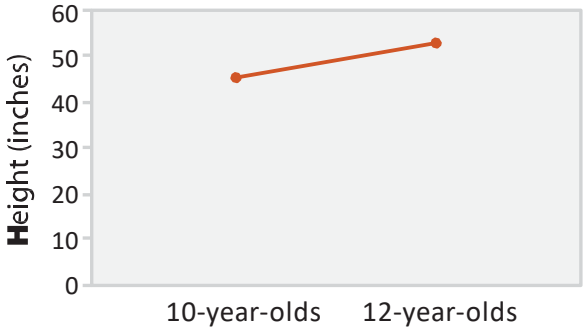
- Similar to scatterplots but allow ordered attribute

Proper Use of Line and Bar Charts





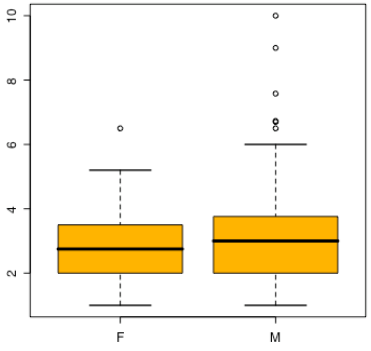


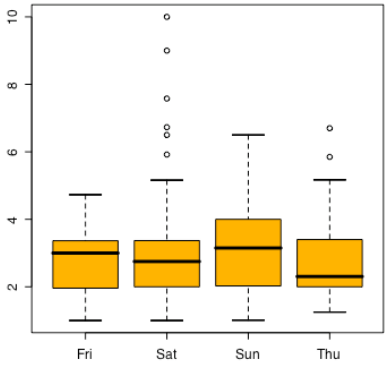


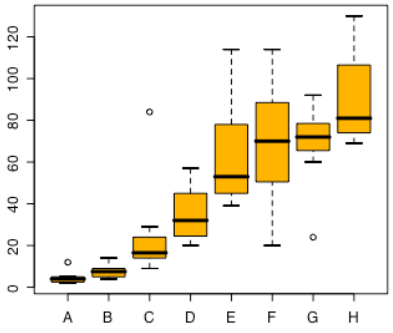
Two Variables - one categorical

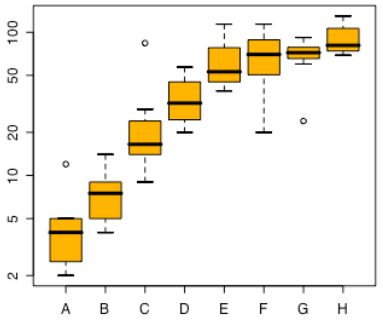
Side by side boxplots are very effective in showing differences in a quantitative variable across factor levels

- tips data
  - do men or women tip better
- orchard sprays
  - measuring potency of various orchard sprays in repelling honeybees



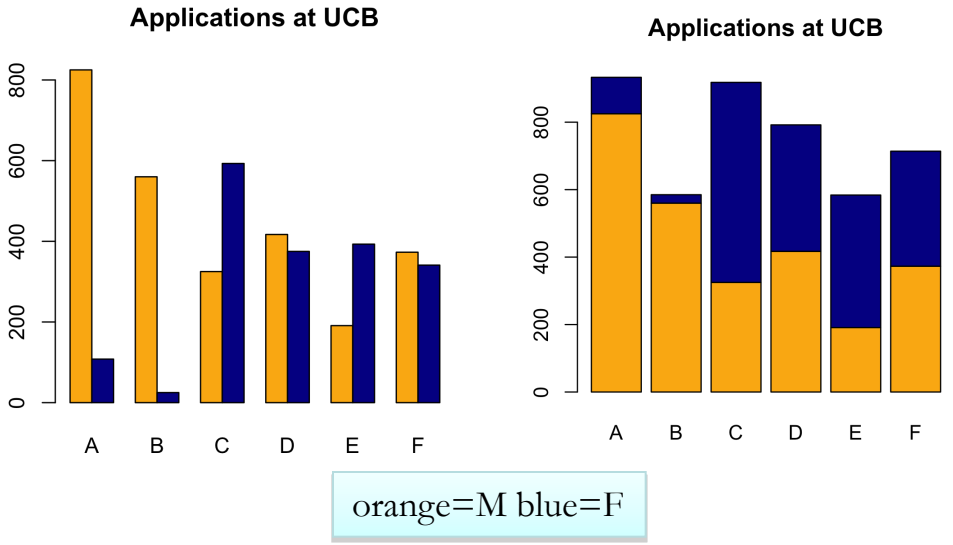




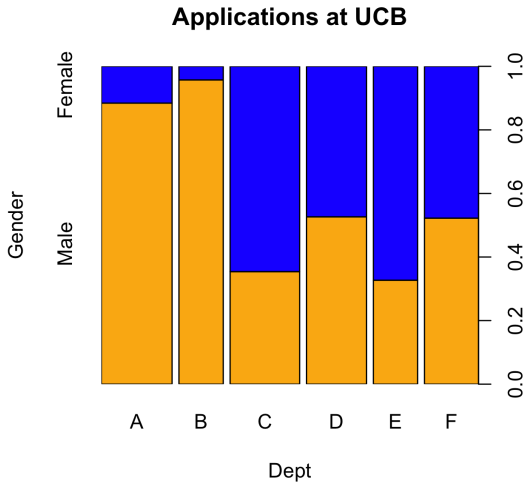


Barcharts and Spineplots

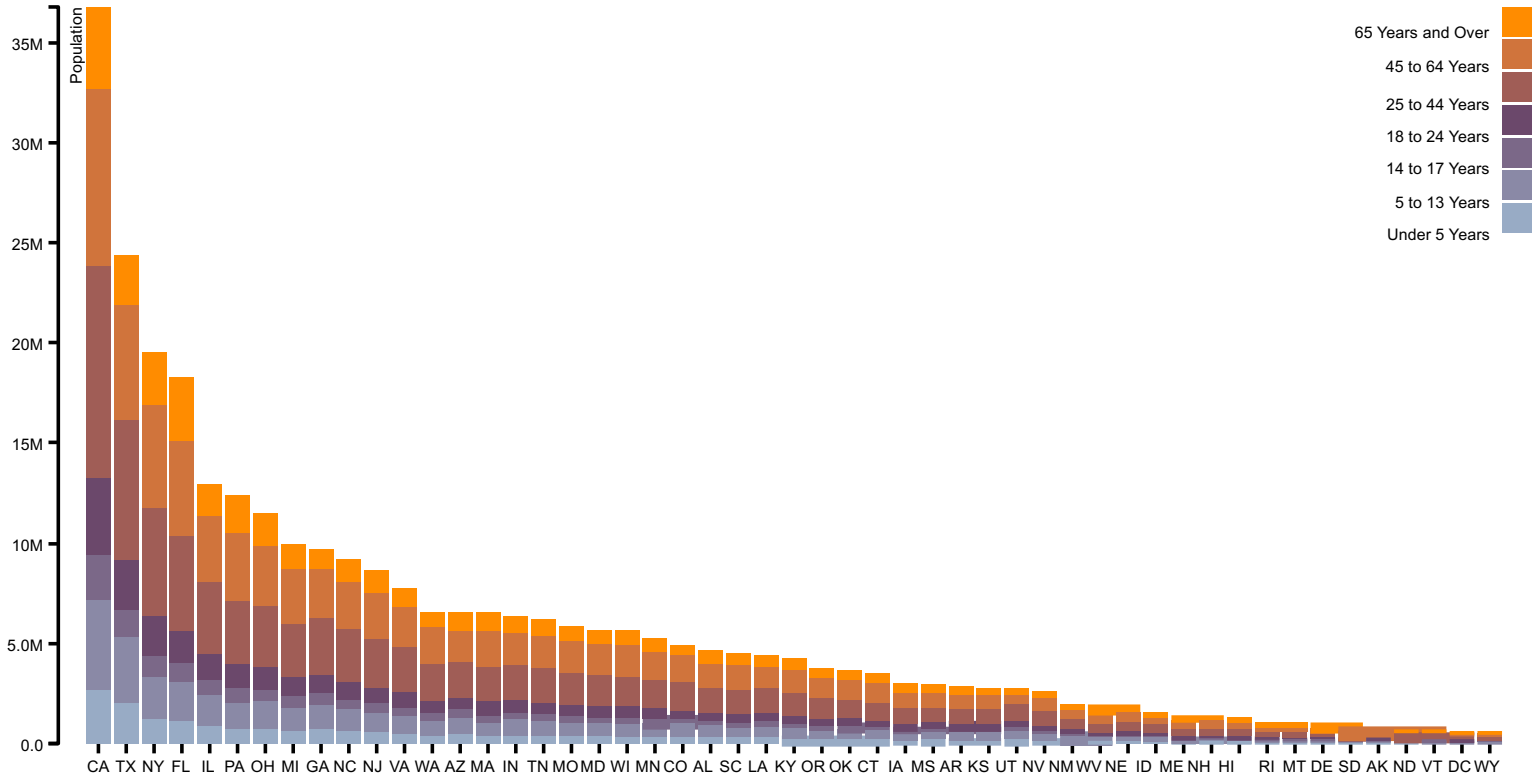
stacked barcharts can be used to compare continuous values across two or more categorical ones.



spineplots show proportions well, but can be hard to interpret

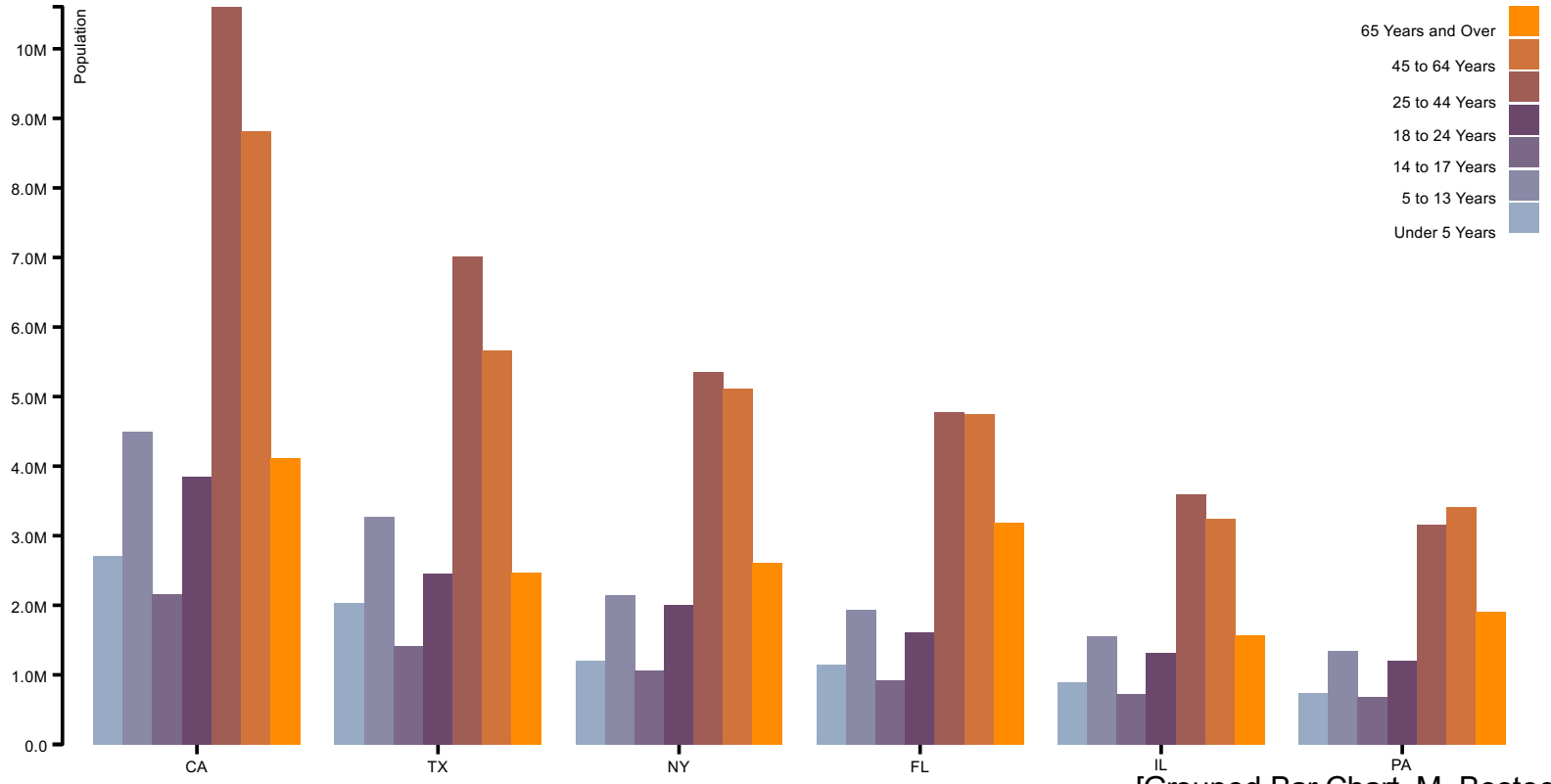


Stacked Bar Charts



[Stacked Bar Chart, M. Bostock, 2017]

Grouped Bar Chart



[Grouped Bar Chart, M. Bostock, 2017]

Stacked Bar Charts

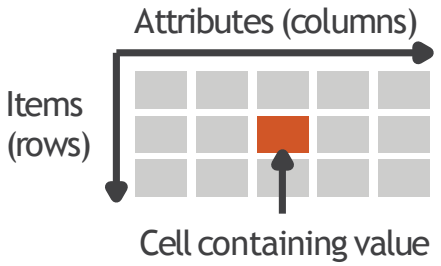
- Data: multidimensional table: one quantitative, two categorical
- Task: lookup values, part-to-whole relationship, trends
- How: line marks: position (both horizontal & vertical), subcomponent line marks: length, color
- Scalability: main axis (hundreds like bar chart), bar classes (<12)

- Orientation: vertical or horizontal (swap how horizontal and vertical position are used.

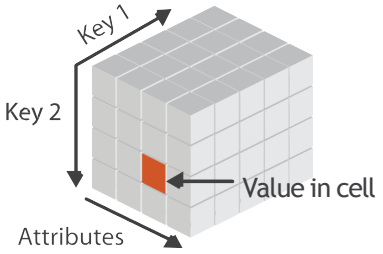
In short

Visualization of Tables

- Items and attributes
- For now, attributes are not known to be positions
- Keys and values
  - key is an independent attribute that is unique and identifies item
  - value tells some aspect of an item
- Keys: categorical/ordinal
- Values: +quantitative
- Levels: unique *values* of categorical or ordered attributes



➔ *Multidimensional Table*



[Munzner (ill. Maguire), 2014]

Arrange Tables

- **Express Values**

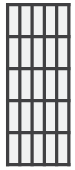
- **Separate, Order, Align Regions**
  - Separate

  - Order

  - Align

    - *1 Key*
      List

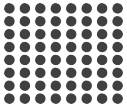
    - *2 Keys*
      Matrix



    - *3 Keys*
      Volume

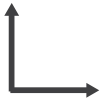
    - *Many Keys*
      Recursive Subdivision

- **Layout Density**
  - Dense



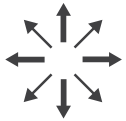
  - Space-Filling

- **Axis Orientation**
  - Rectilinear



  - Parallel

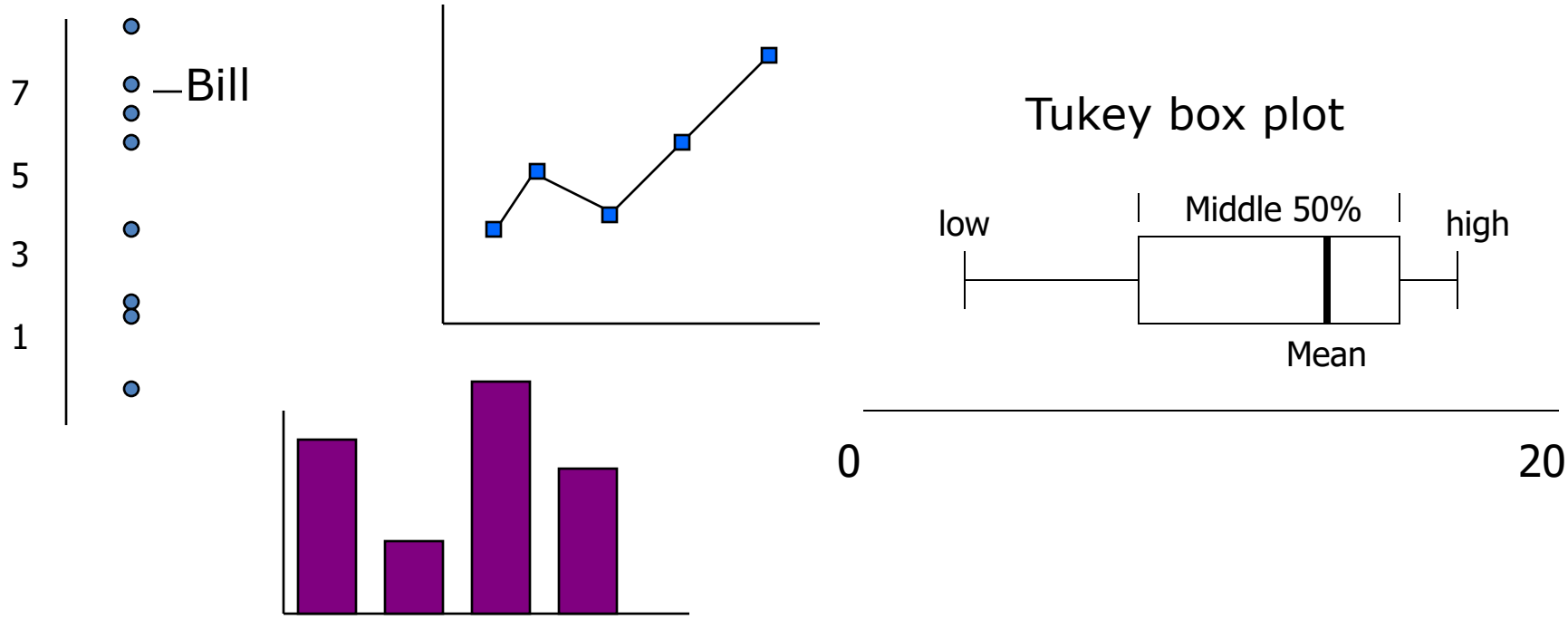
  - Radial



[Munzner (ill. Maguire), 2014]

Univariate Data

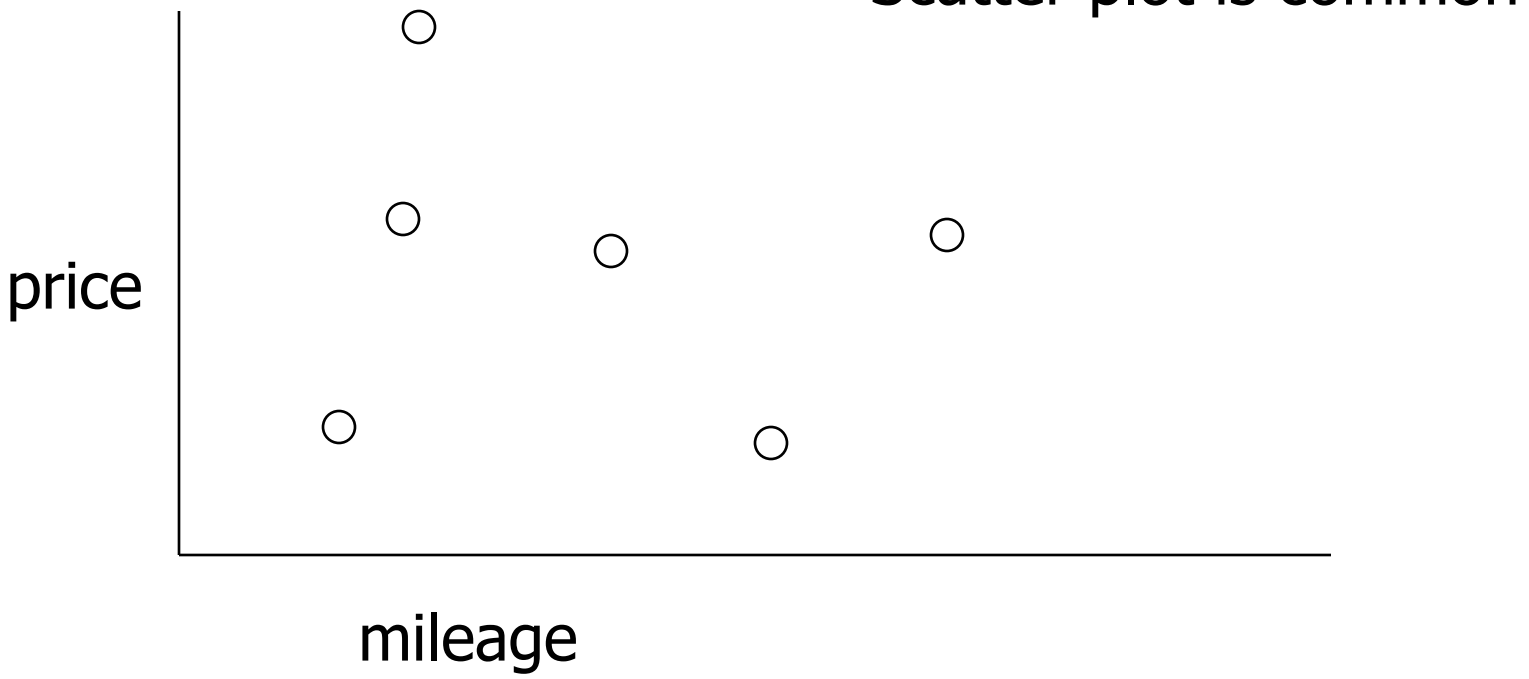
Representations



Bivariate Data

Representations

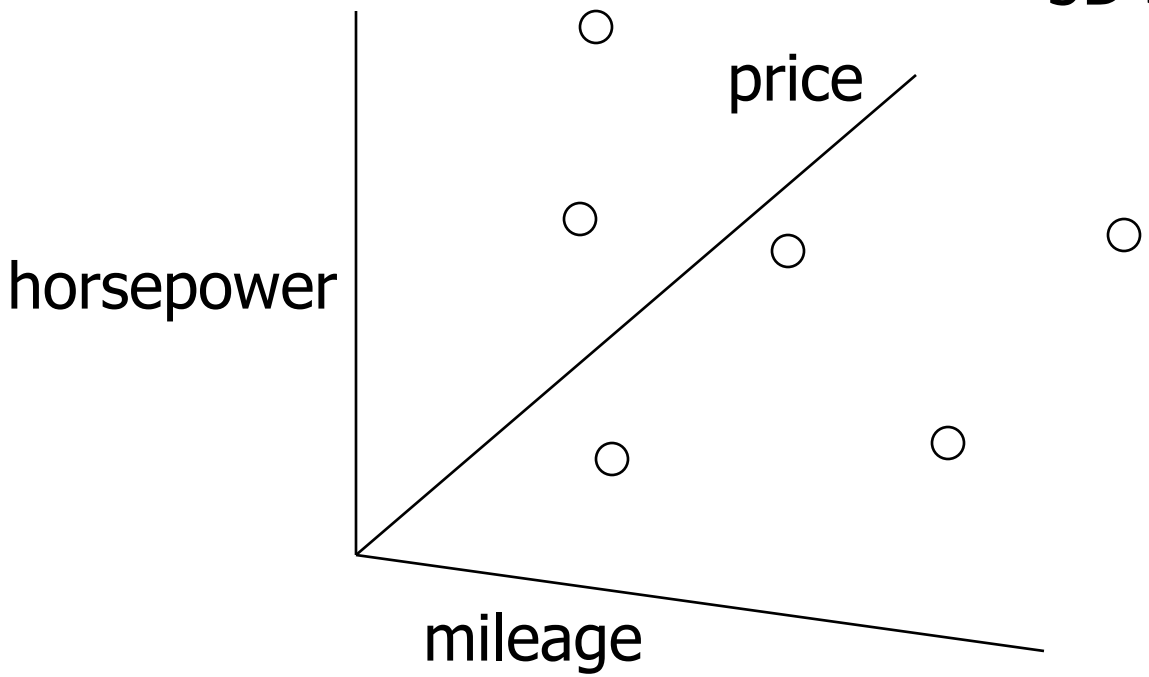
Scatter plot is common



Trivariate Data

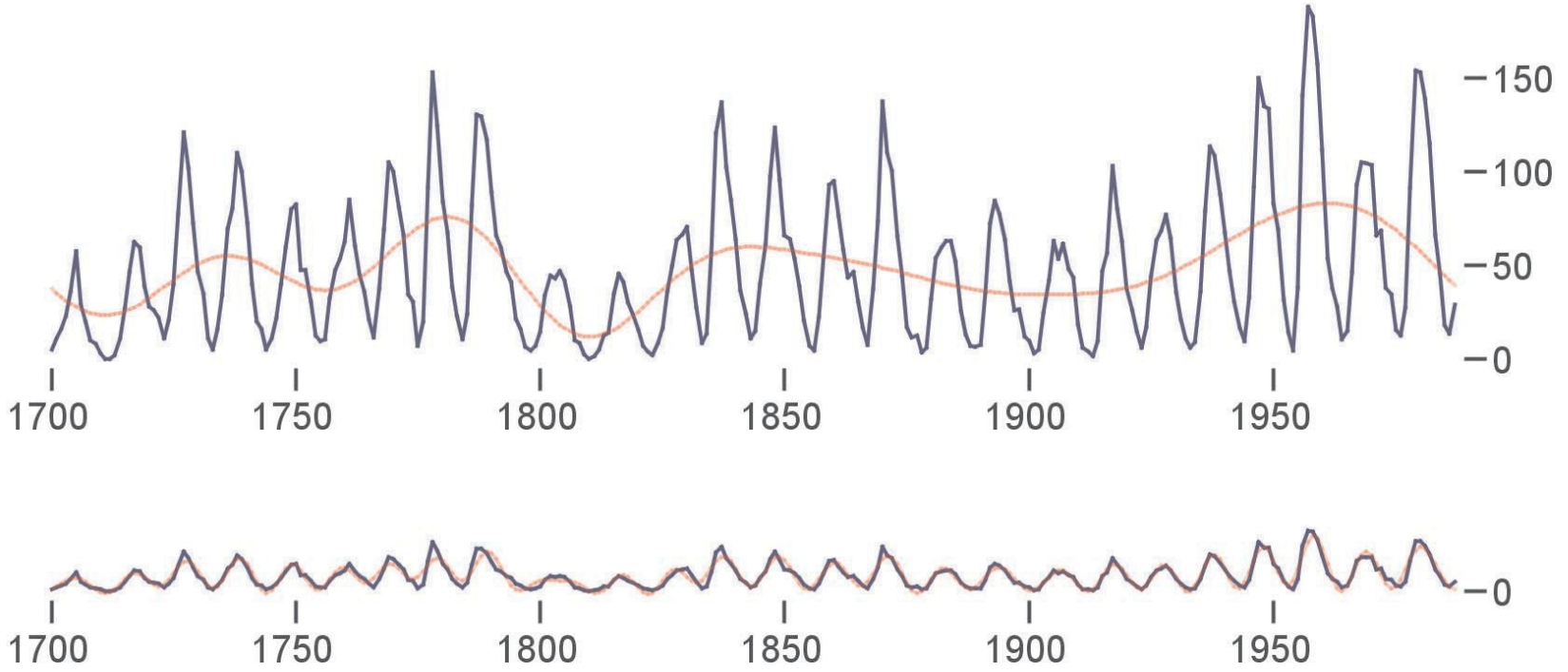
Representations

3D scatter plot is possible

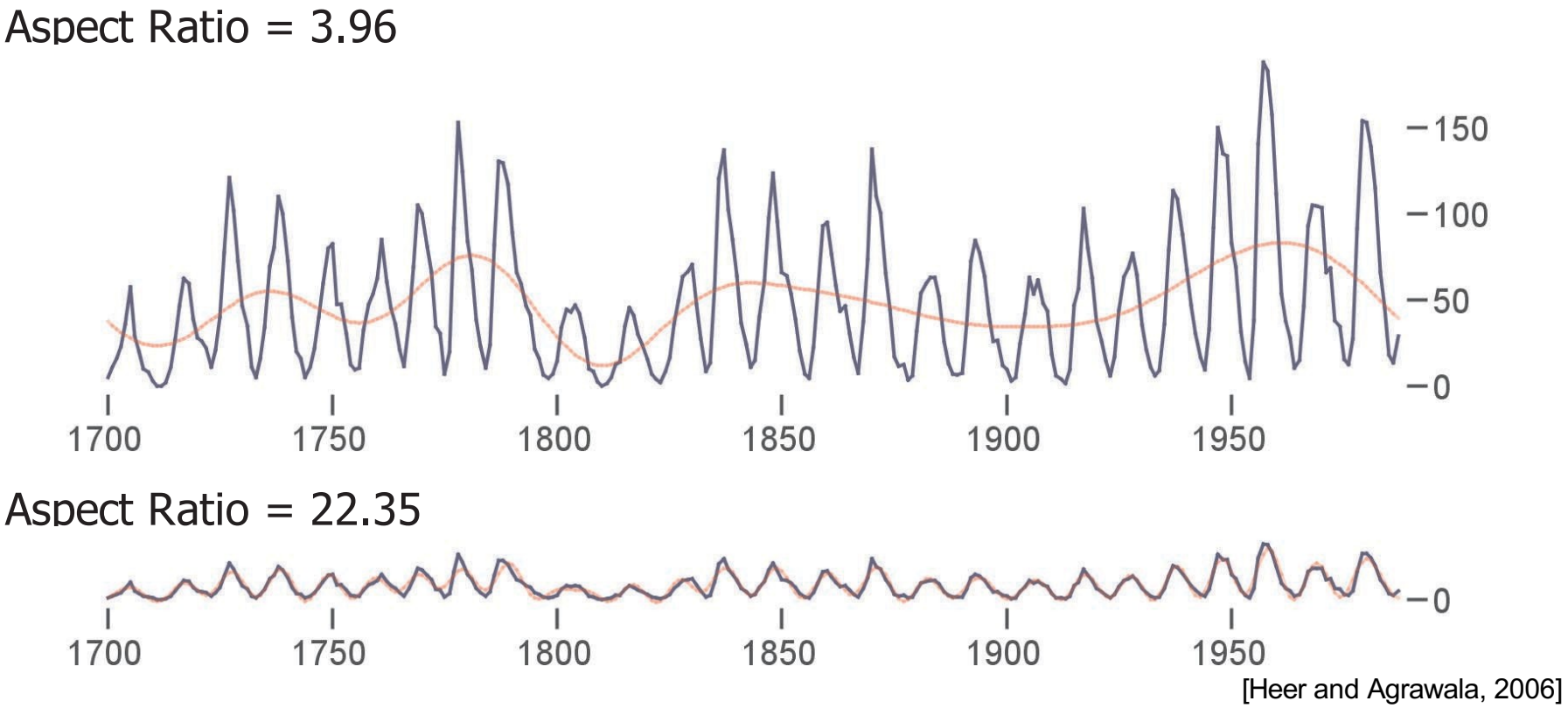


Some more points

Anything in common? (in the trends below)



Multiscale Banking

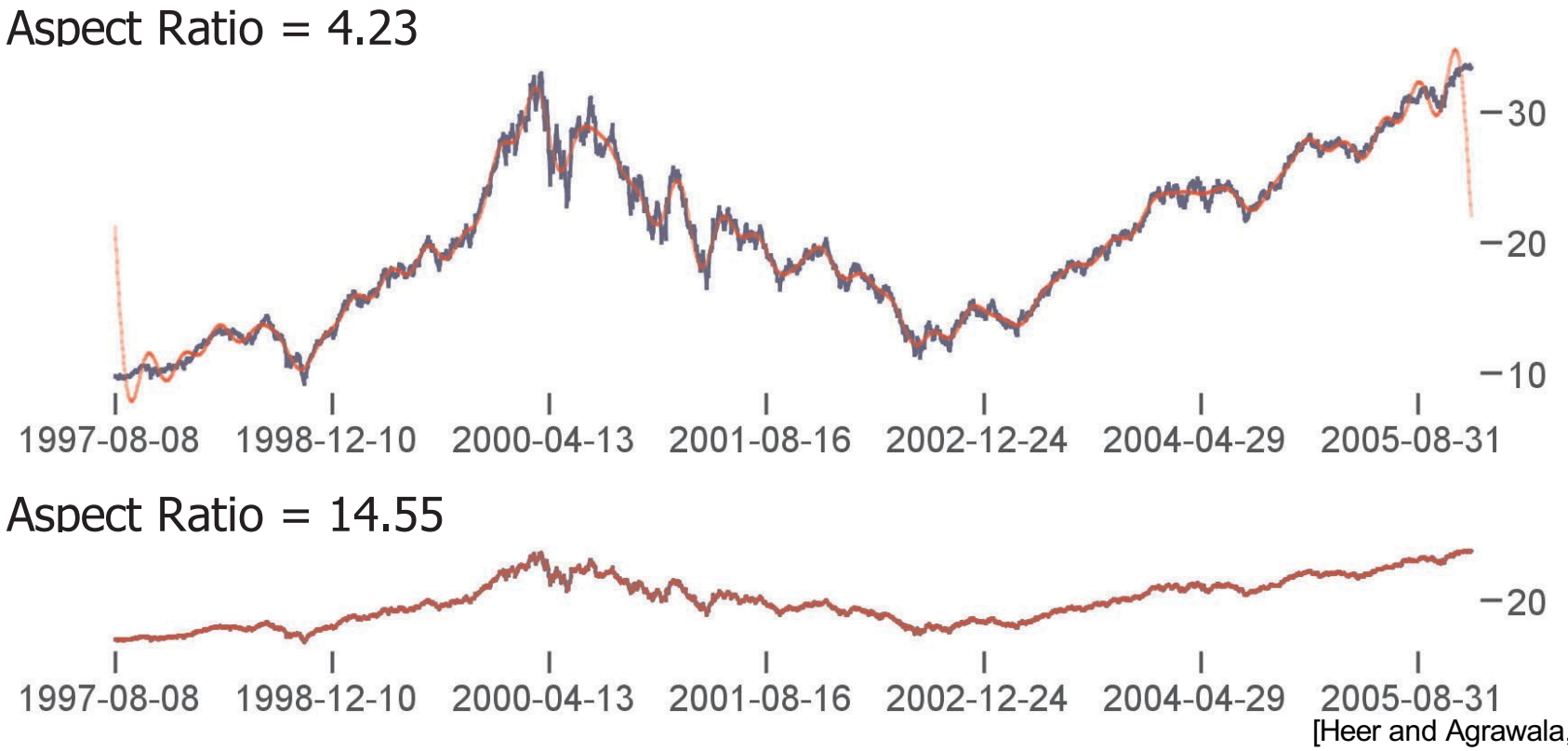


[Heer and Agrawala, 2006]

IMPORTANT: Aspect Ratio

- Trends in line charts are more apparent because we are using angle as a channel
- Perception of angle (and the relative difference between angles) is important
- Initial experiments found people best judge differences in slope when angles are around 45 degrees (Cleveland et al., 1988, 1993)

Multiscale Banking



[Heer and Agrawala, 2006]

Task

Study different data visualization plots and diagrams at

www.datavizcatalogue.com