Plotting multivariate data

Statistics 407
ISU

Multivariate data and graphics

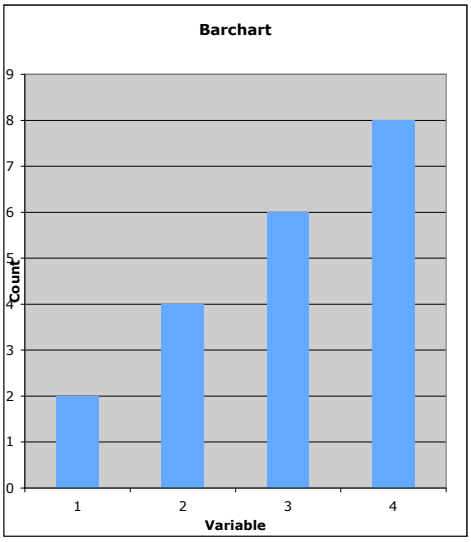
With multivariate data we want to understand the associations between multiple variables, which might be considered to be understanding the shape of the data in high-dimensional space.

Graphics are used to explore the data, and also to diagnose models.

Typically start with plots of 1, then 2, then more variables.

One variable - categorical

Eg Examining the number in each level of one of the label variables, prior to doing classification or MANOVA



* What do you think the counts are for each of the categories?
* Naomi Robbins (2004) "Creating More Effective Graphs" Wiley

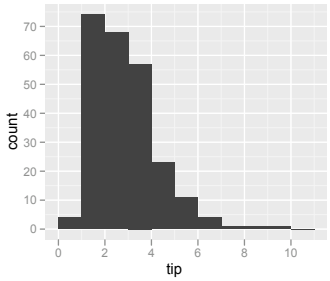
Counts are 2, 4, 6, 8

PLEASE DON'T USE 3D UNNECESSARILY

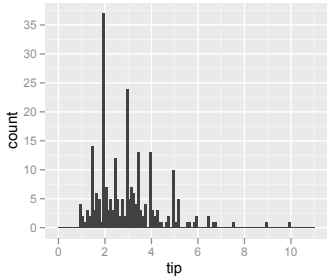
One variable - real

* Histogram: bin the values, and represent counts as bars. Use different bin sizes.
* Density: Like a smoothed histogram, estimate the density, represent as a curve.
* Dotplot: 1D scatterplot, values represented by dots.
* Boxplot: Representation of the 5 number summary, min/max, Q1/Q3, median. Good for comparing distrbutions of multiple categories.

One variable - real

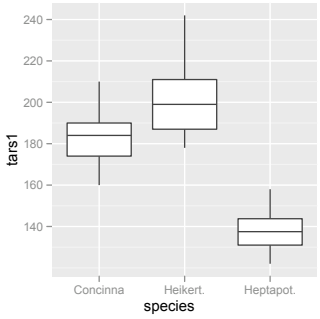


What do you learn about tips?

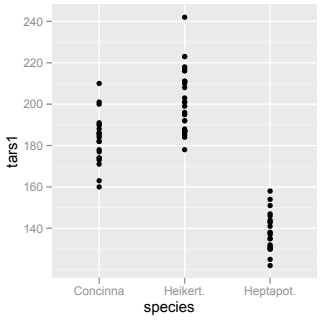


What else do you learn about tips?

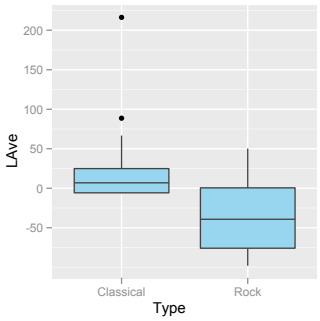
One real + one categorical



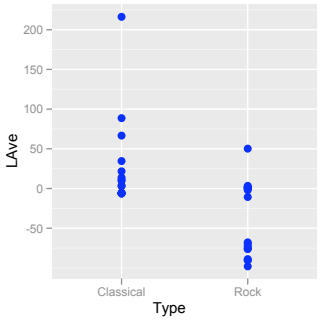
How do the species differ in the tars1 values?



One real + one categorical

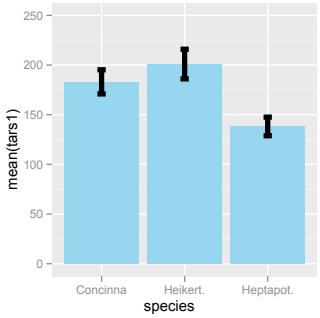


How does the type of music differ?



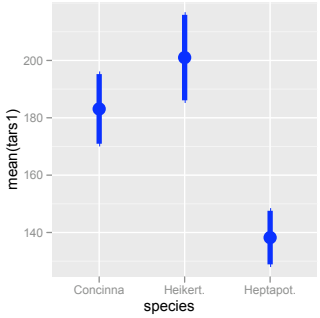
Do you learn anything different from the dotplot than the boxplot?

Plotting Means



What's wrong with this plot?

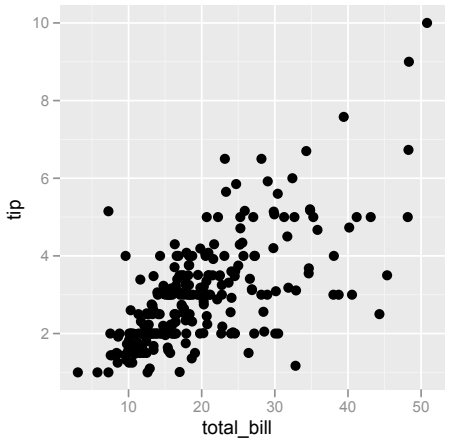
Means are points estimates, use points to represent them.



What do we learn about differences between species?

Two variables - real

* Scatterplot

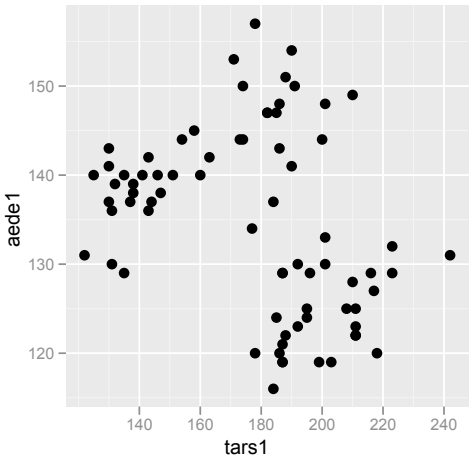


How is tip related to bill?
How would you expect tip to be related to bill?

Does aspect ratio matter?
Which variable is the response and which is the explanatory?

Two variables - real

* Scatterplot



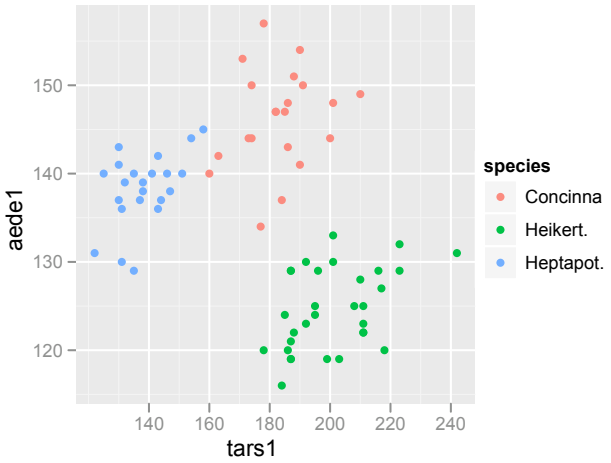
How is aede1 related to tars1?

Does aspect ratio matter?
Which variable is the response and which is the explanatory?

For scatterplots of multivariate data use a SQUARE aspect ratio!

Two real + one categorical

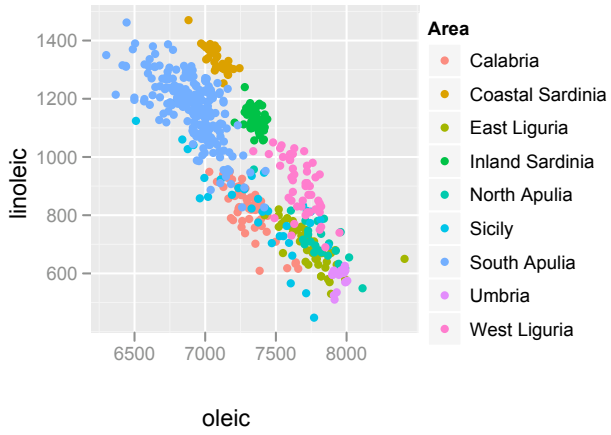
* Scatterplot



How is aede1 related to tars1, by species?

Two real + one categorical

* Scatterplot



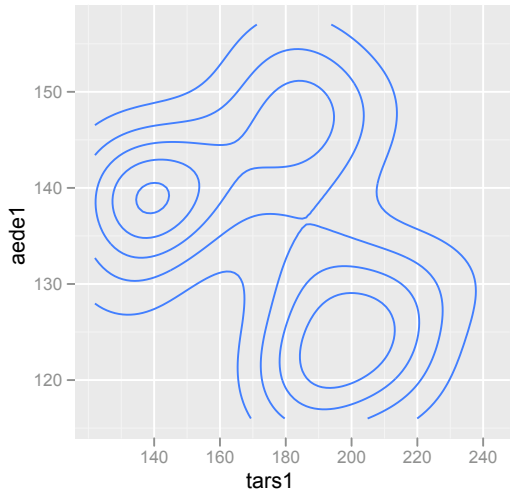
How is linoleic acid related to oleic acid, by area?

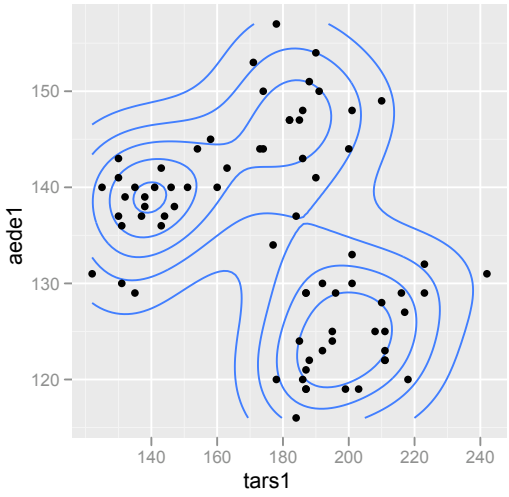
How many colors is too many? Generally only use 3-4 colors in a plot, only 3-4 categories.

Two variables - real

* 2D density

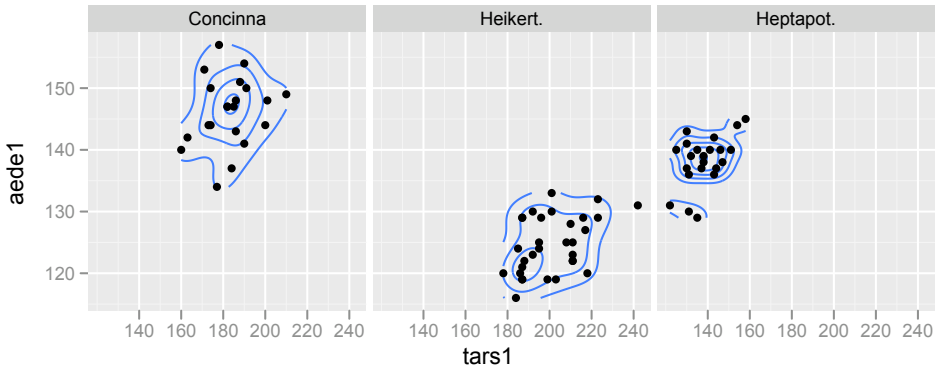
How is aede1 related to tars1?





Two real + one categorical

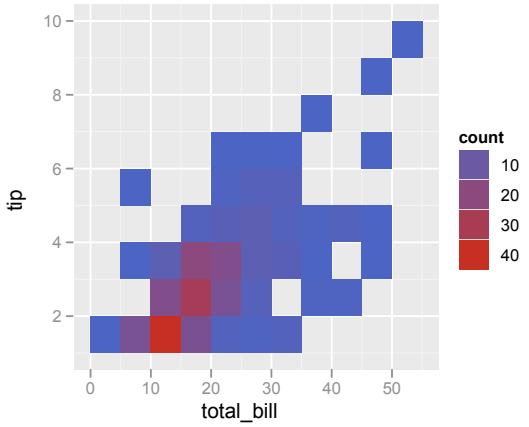
* Scatterplot + contour, facetted by species



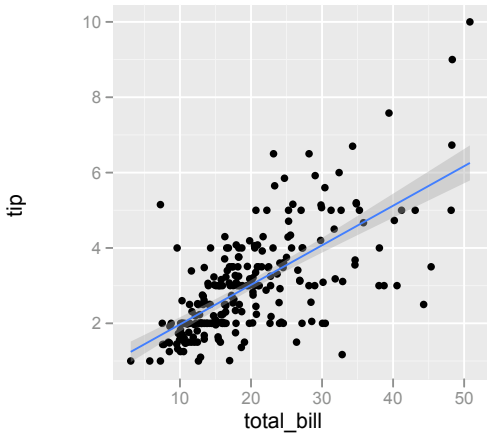
How is aede1 related to tars1, by species?

Two variables - real

* 2D histogram, scatterplot + linear model

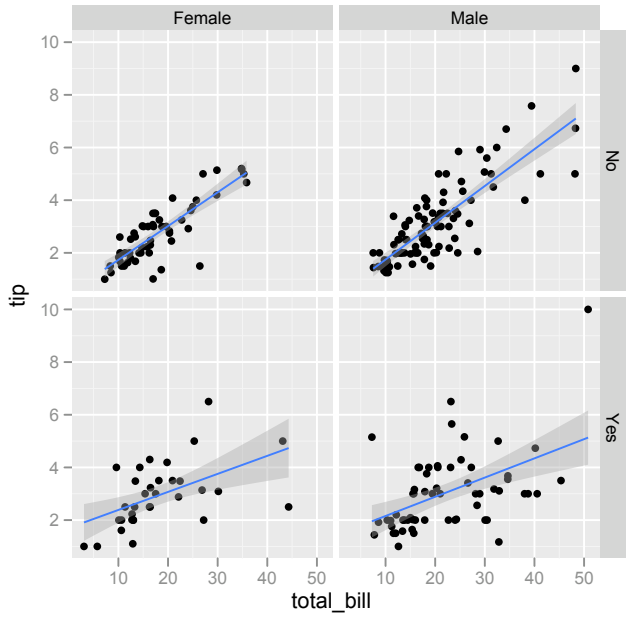


How is tip related to bill?



Two real + one categorical

* Scatterplot + linear model, facetted by sex of the bill payer, and smoking status.



How is tip related to total bill, by sex and smoker?

BEYOND 2D.....

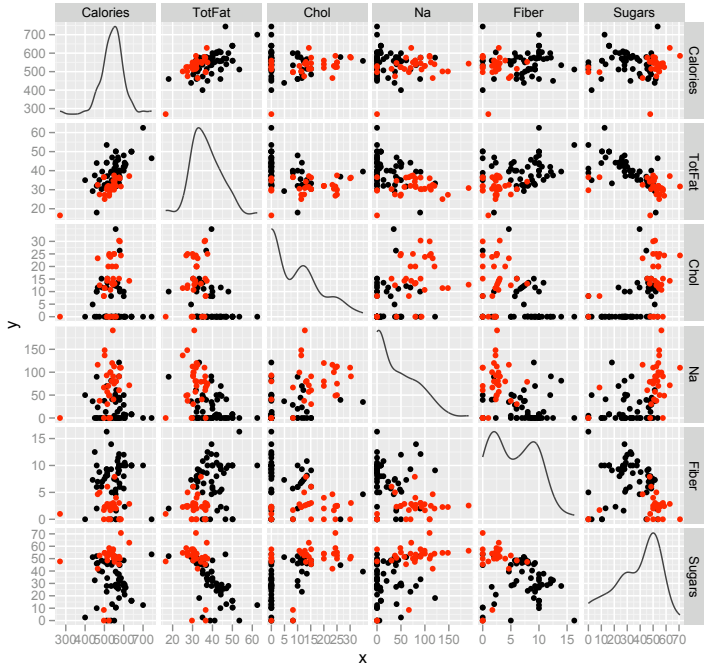
Scatterplot matrix

Correlation matrix.

How do correlations differ by type of chocolate?

```
> cor(chocolates[,c(5,7,9,10,12,13)])
         Calories TotFat   Chol     Na Fiber Sugars
Calories    1.000   0.61 -0.054 -0.057  0.14 -0.092
TotFat      0.611   1.00 -0.410 -0.429  0.39 -0.654
Chol       -0.054  -0.41  1.000  0.491 -0.40  0.540
Na         -0.057  -0.43  0.491  1.000 -0.36  0.467
Fiber       0.140   0.39 -0.397 -0.364  1.00 -0.261
Sugars     -0.092  -0.65  0.540  0.467 -0.26  1.000
> cor(chocolates[chocolates$Type=="Milk",c(5,7,9,10,12,13)])
         Calories  TotFat Chol   Na   Fiber Sugars
Calories    1.000  0.7359 0.37 0.32  0.0130   0.22
TotFat      0.736  1.0000 0.22 0.11 -0.0097  -0.21
Chol        0.368  0.2211 1.00 0.40  0.1209   0.33
Na          0.323  0.1060 0.40 1.00  0.1510   0.44
Fiber       0.013 -0.0097 0.12 0.15  1.0000   0.20
Sugars      0.223 -0.2141 0.33 0.44  0.2039   1.00
> cor(chocolates[chocolates$Type=="Dark",c(5,7,9,10,12,13)])
         Calories TotFat  Chol     Na   Fiber  Sugars
Calories    1.000   0.60 -0.15 -0.117  0.0419 -0.1235
TotFat      0.595   1.00 -0.36 -0.299  0.1177 -0.6685
Chol       -0.153  -0.36  1.00  0.103 -0.2233  0.4303
Na         -0.117  -0.30  0.10  1.000 -0.0538  0.0793
Fiber       0.042   0.12 -0.22 -0.054  1.0000  0.0012
Sugars     -0.124  -0.67  0.43  0.079  0.0012  1.0000
```

Scatterplot matrix



Red=milk
Black=dark

All variables plotted pairwise, like correlation matrix.

What do you learn?

Scatterplot matrix

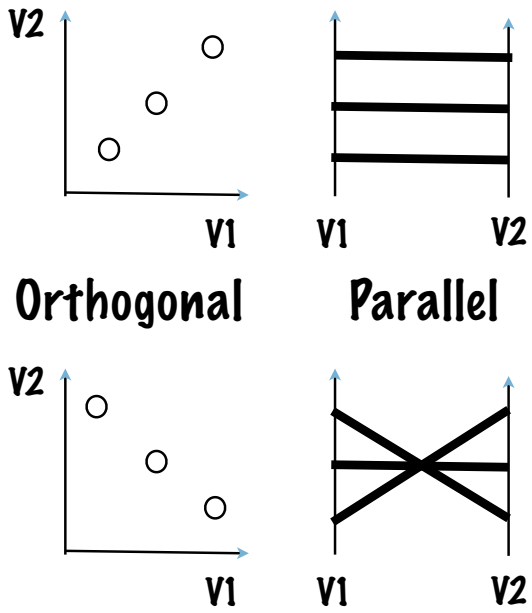
V1 density	V2 vs V1	V3 vs V1
V1 vs V2	V2 density	V3 vs V2
V1 vs V3	V2 vs V3	V3 density

GENERAL

Pairwise dependence only.

A cross-check on the correlation matrix, whether correlation is a good summary or NOT.

Parallel coordinates



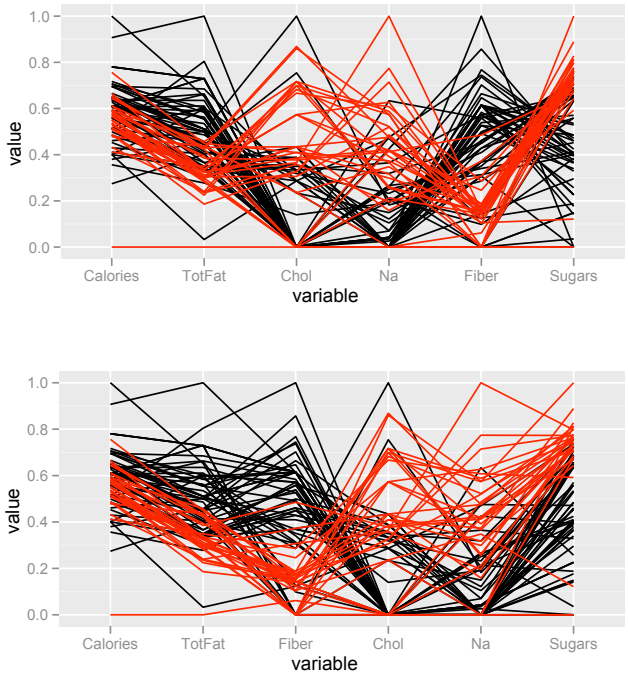
GENERAL

Orthogonal axes become parallel.

Can lay out many variables on a page. Need to learn a new set of patterns of structure.

Re-ordering of axes is important.

Parallel coordinates



Red=milk
Black=dark

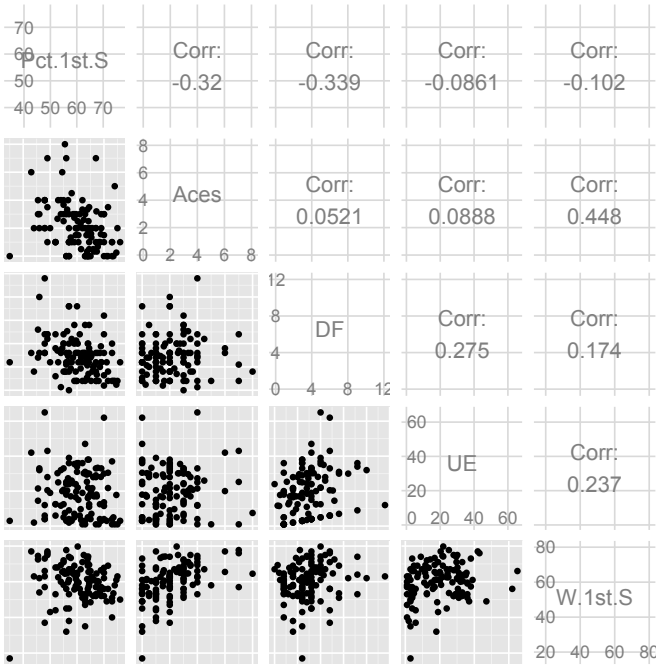
Re-ordered so that variables with high values for dark chocolate are first.

What's the difference between milk and dark chocolate?

Tennis data

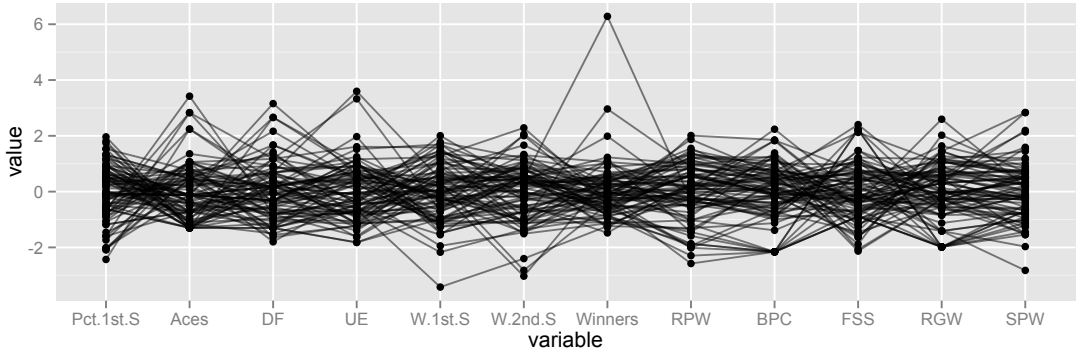
- 2012 Australian Open tennis tournament statistics for women
- % 1st serves in, Aces, Double faults, Unforced errors, Win on 1st serve %, Win on 2nd serve %, Winners, Receiving points won, Break point conversions, Fastest serve speed, Return games won, Server points won (12 variables)

Scatterplot matrix

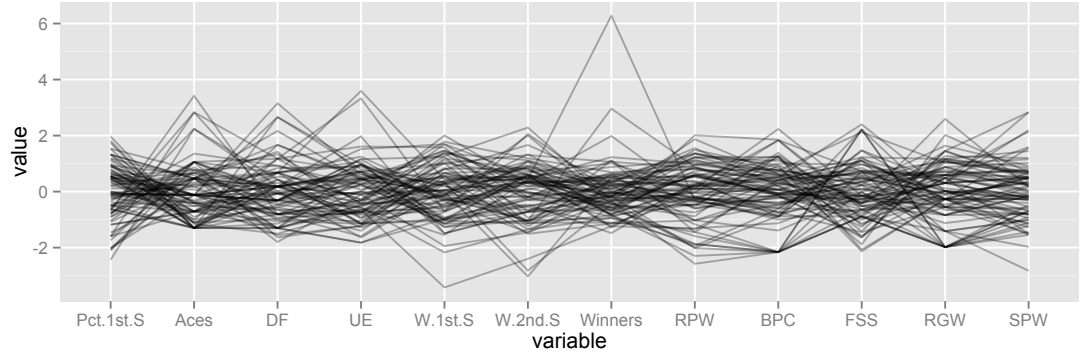


- Weak associations
- Outliers
- Can't plot many variables

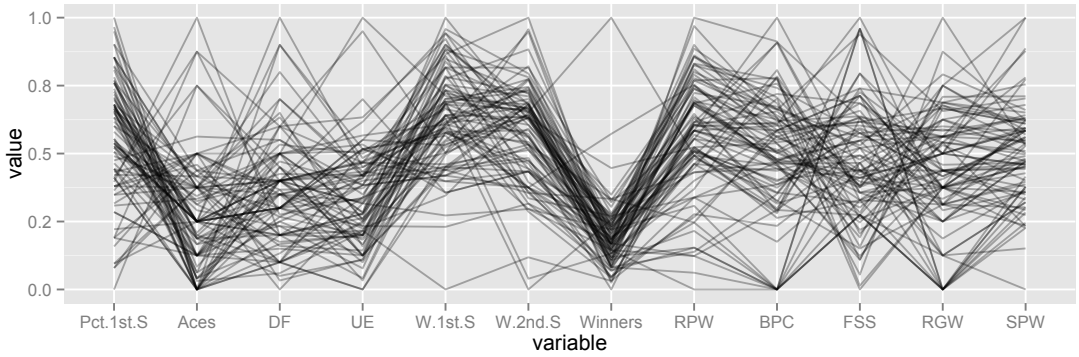
Parallel coordinates



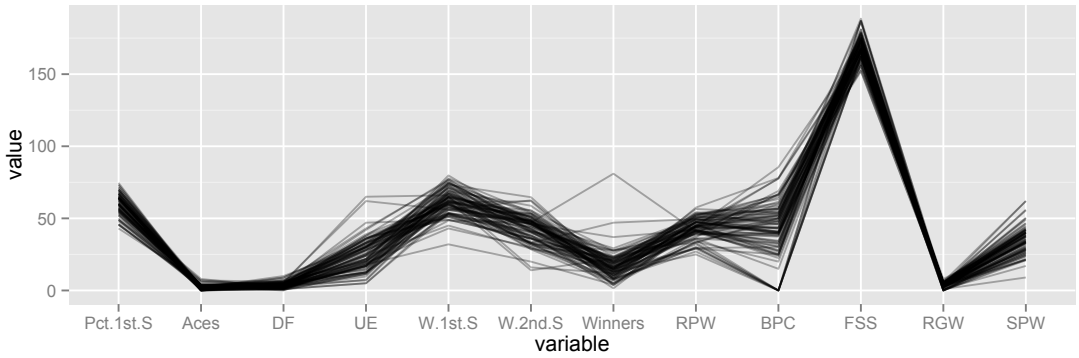
* Plot a lot more variables
* Some outliers
* Some weak associations can be seen, both positive and negative



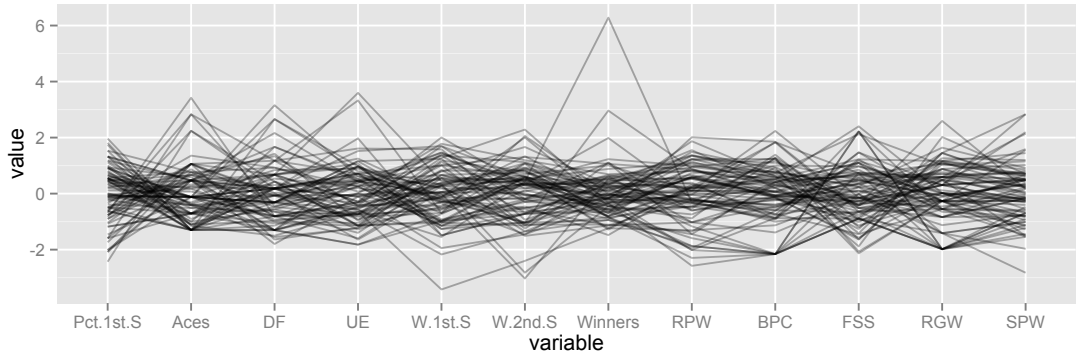
* Scale matters: mean/sd, 0/1, individual/global
* Enables correlation to be seen better, and outliers



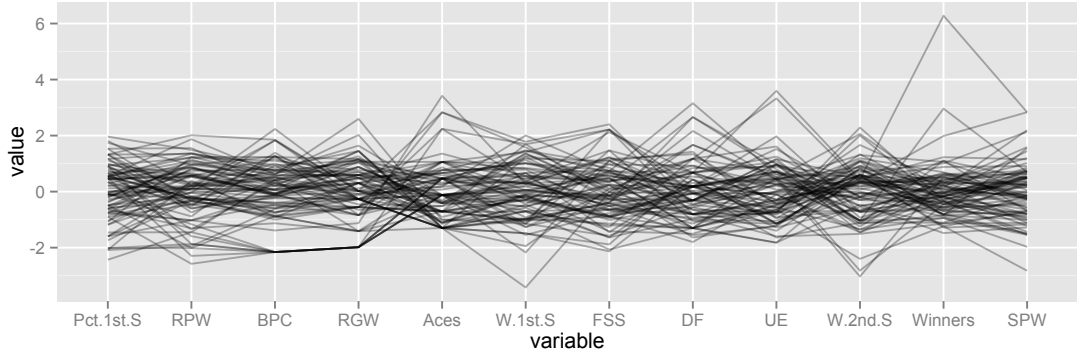
- Scale matters: mean/sd, 0/1, individual/global
- Emphasizes the univariate distributions



- Scale matters: mean/sd, 0/1, individual/global

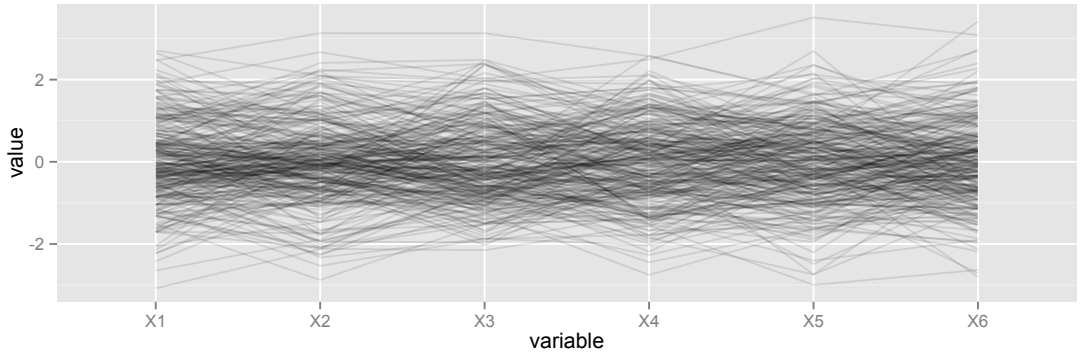


* Order matters: place variables that are highly correlated close to each other



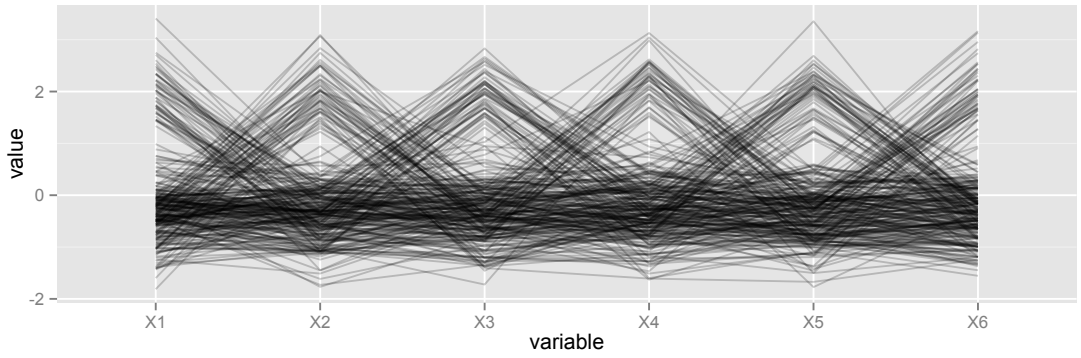
* Order matters: place variables that are highly correlated close to each other
* Less line crossing, easier to digest positive correlation, and then negative correlation

Normal data



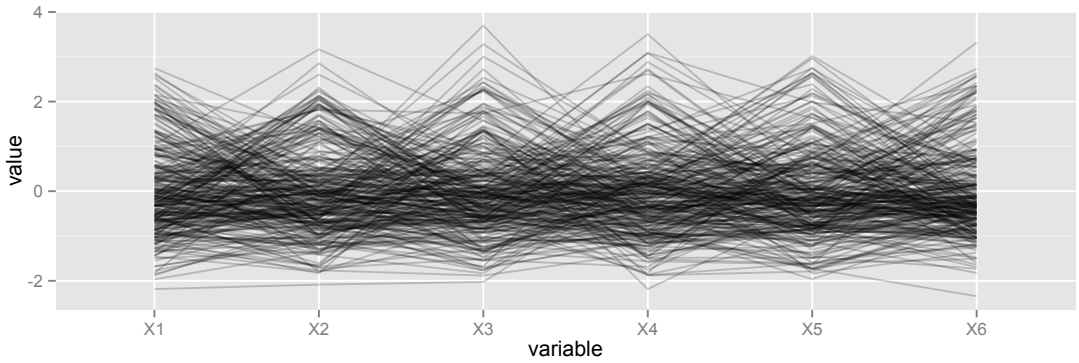
- Nothing interesting! All a little moderate correlation.
- Modelling is going to be easy!

Clustered data



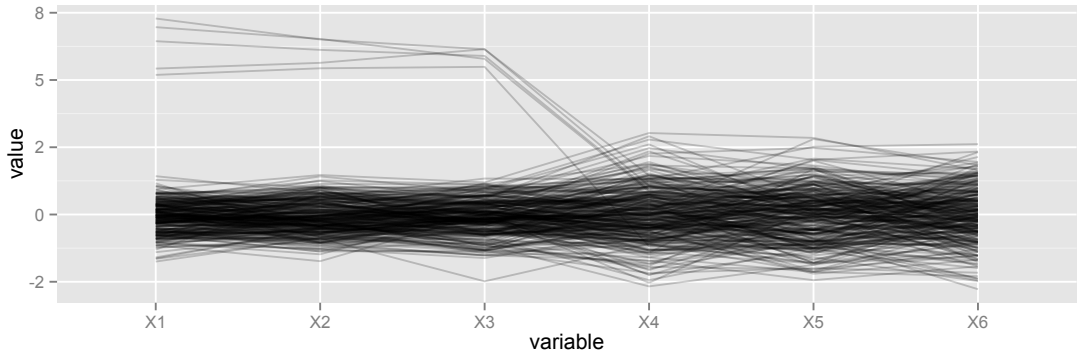
- See the criss-crossing, gaps between lines.
- Will need to extract the clusters before doing any other modeling, otherwise pretty regular data

(Less) Clustered data



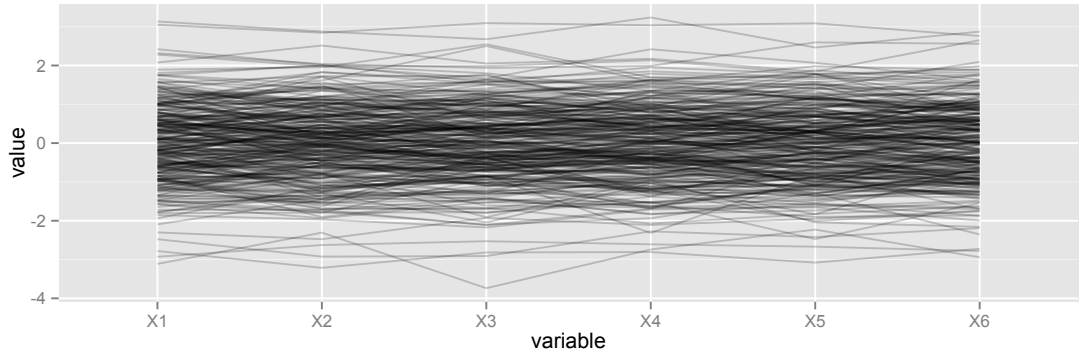
* Still see the criss-crossing, gaps between lines, but less prominent.
* Will need to deal with the multi-modality

Outliers in the data



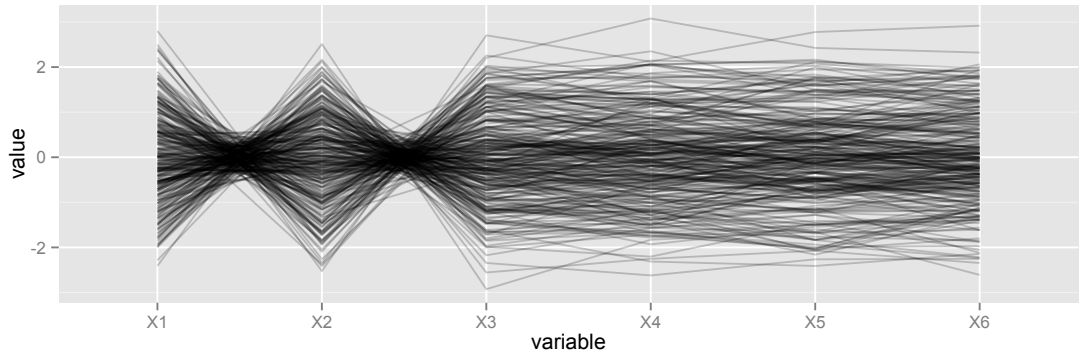
* Small group of observations that are outliers on X1-X3.
* Need to do something with these cases, remove with justification, or fix

Strong association



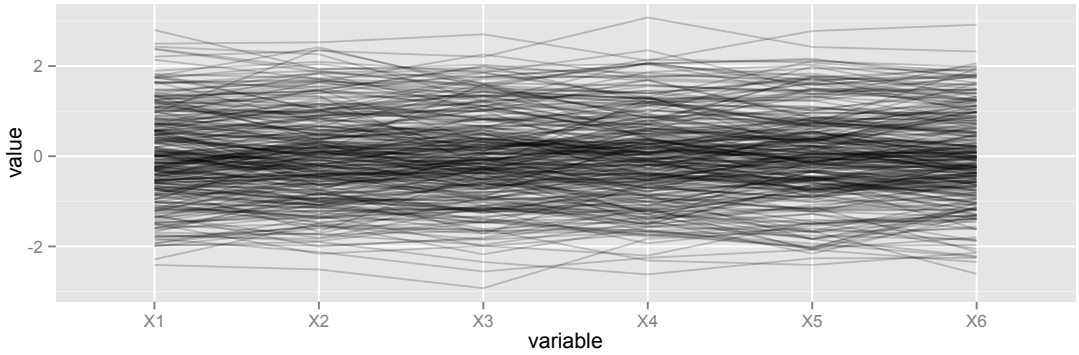
* Very flat lines indicate strong positive association between all variables.

Strong negative association

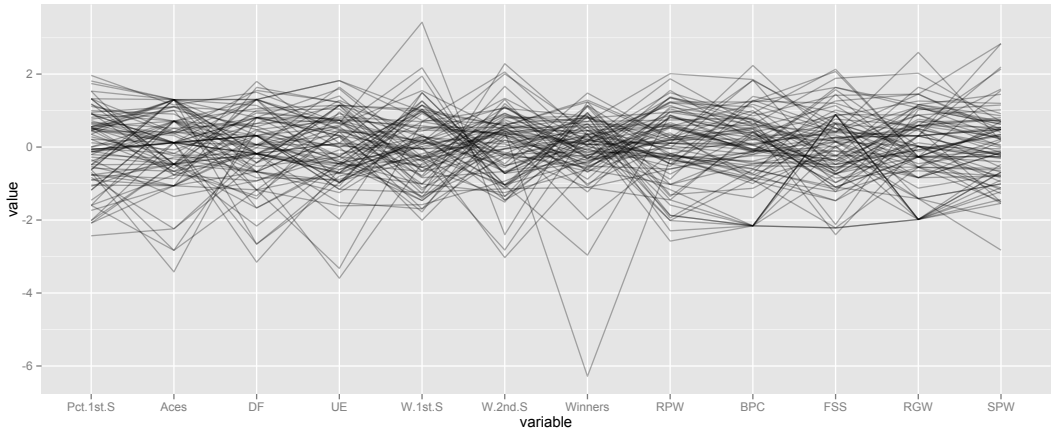


* Crossed lines in first three variables indicate X2 is strongly negatively correlated with other vars.

Strong negative association - fixed

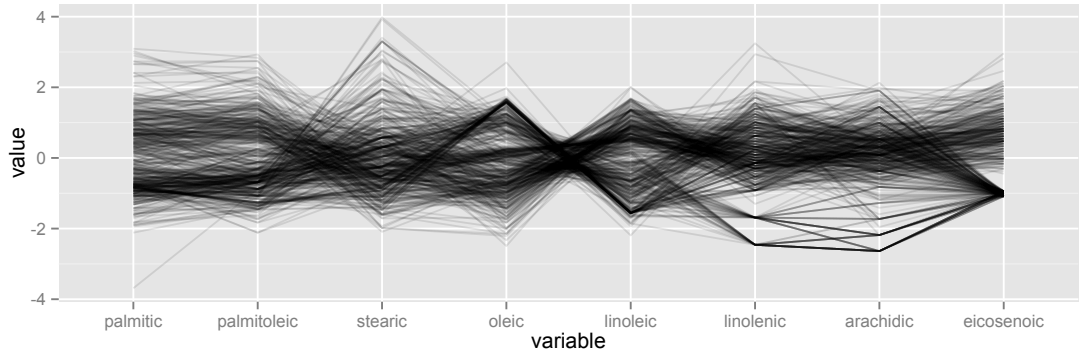


* X2 is multiplied by -1, then it is positively associated with other variables.



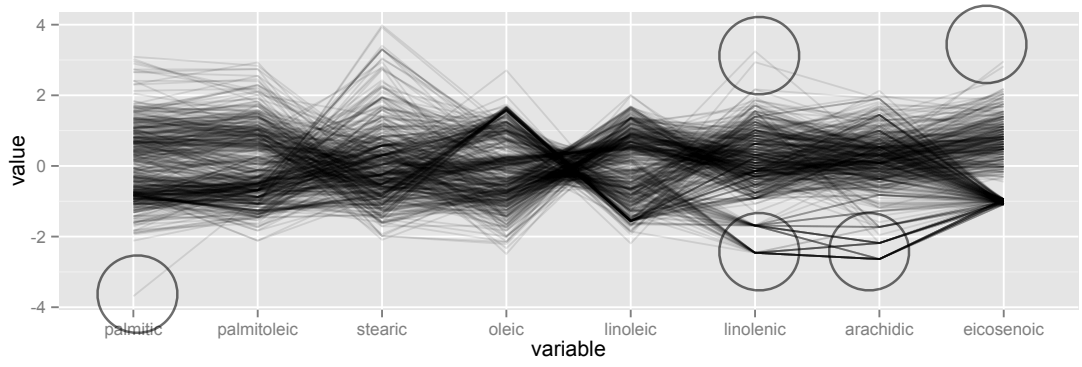
* Tennis data again: Aces, DF, UE, W.1st.S, Winners, FSS reversed (multiplied by -1)
* Mostly now the outliers are visible, but not so much multivariate outliers, mostly only in one or few variables

Another data set: olive oils



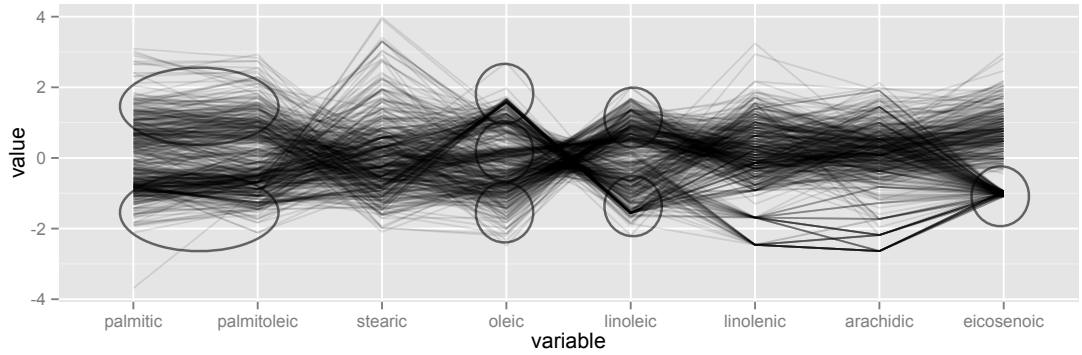
* Outliers, clustering, some positive and some negative associations, some discreteness

Another data set: olive oils



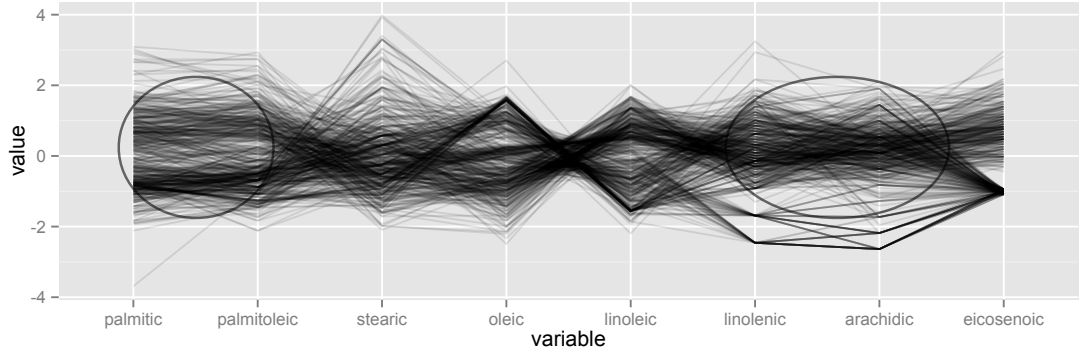
* Outliers, clustering, some positive and some negative associations, some discreteness

Another data set: olive oils



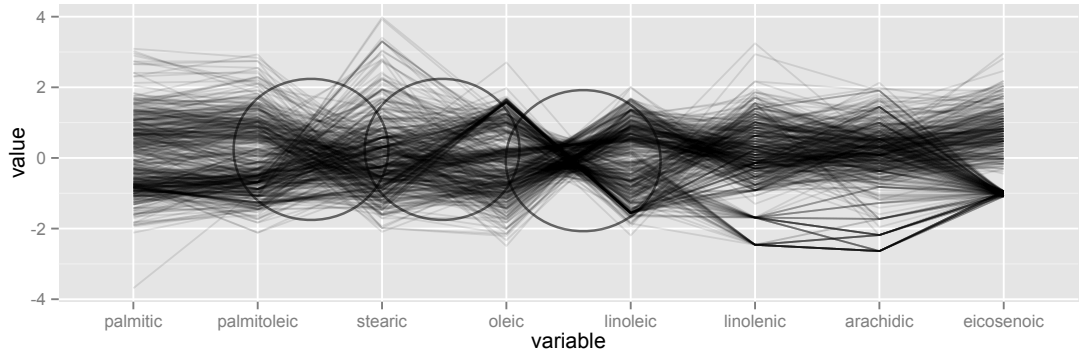
* Outliers, clustering, some positive and some negative associations, some discreteness

Another data set: olive oils



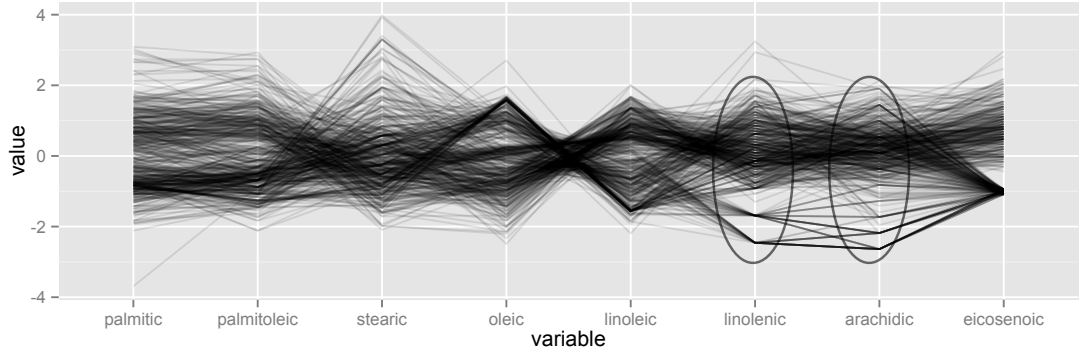
* Outliers, clustering, some positive and some negative associations, some discreteness

Another data set: olive oils



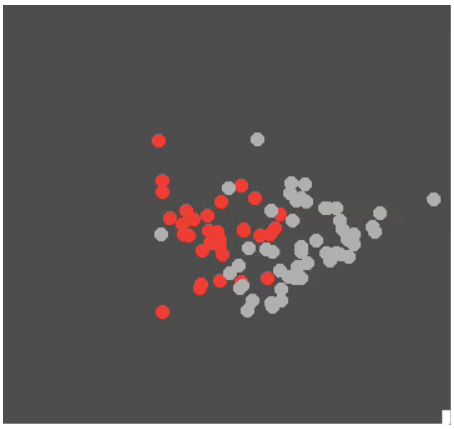
* Outliers, clustering, some positive and some negative associations, some discreteness

Another data set: olive oils



* Outliers, clustering, some positive and some negative associations, some discreteness

Tours



Red=milk
Grey=dark

Are the two clusters different?

How?

Are there any outliers? And points from one group that seem to belong to the other group?

Tours

A tour is a movie of low-dimensional projections of high-dimensional space.

Let

$$\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2] = \begin{bmatrix} a_{11} & a_{12} \\ \vdots & \\ a_{p1} & a_{p2} \end{bmatrix}$$

be a 2-dimensional projection matrix, where $\mathbf{a}_1$ and $\mathbf{a}_2$ are both of length 1, and orthogonal to each other, then for data matrix $\mathbf{X}_{n\times p}$

$$\mathbf{XA} = \begin{bmatrix} a_{11}X_{11} + a_{21}X_{12} + \ldots + a_{p1}X_{1p} & a_{12}X_{11} + a_{22}X_{12} + \ldots + a_{p2}X_{1p} \\ \vdots & \vdots \\ a_{11}X_{n1} + a_{21}X_{n2} + \ldots + a_{p1}X_{np} & a_{12}X_{n1} + a_{22}X_{n2} + \ldots + a_{p2}X_{np} \end{bmatrix}_{n\times 2}$$

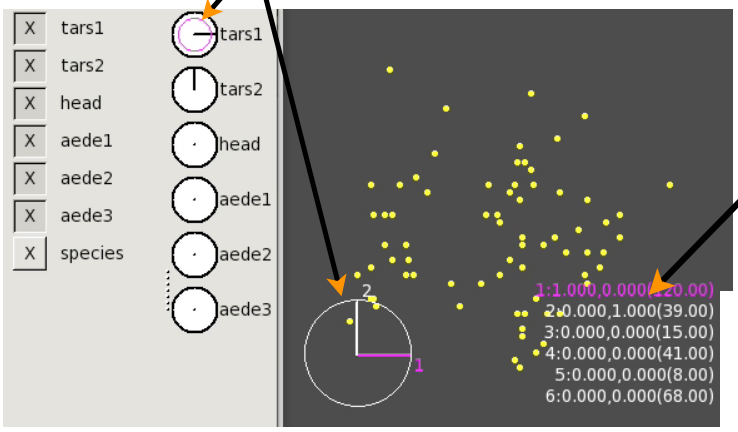
is a 2-dimensional projection of the data.

The values (call them coefficients) in the projection matrix are varied producing the views in the movie.

Tours

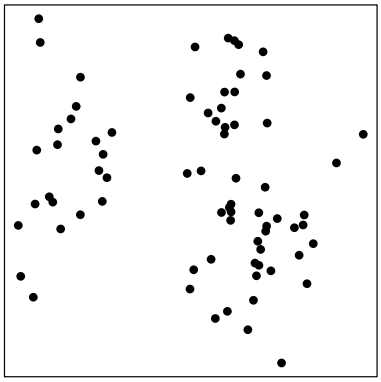
Visual representation of coefficients (axes)

Projection coefficients and scaling

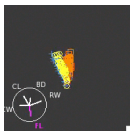


$$\mathbf{A} = \begin{bmatrix} 0.548 & -0.152 \\ -0.152 & 0.754 \\ 0.185 & 0.363 \\ 0.031 & 0.463 \\ 0.793 & 0.116 \\ -0.110 & -0.222 \end{bmatrix} \text{ scaling for each variable } \begin{bmatrix} 0.548/120 & -0.152/120 \\ -0.152/39 & 0.754/39 \\ 0.185/15 & 0.363/15 \\ 0.031/41 & 0.463/41 \\ 0.793/8 & 0.116/8 \\ -0.110/68 & -0.222/68 \end{bmatrix}$$

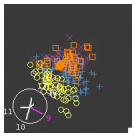
Multiply the data matrix by this to get the particular view of the data.



Types of Tours

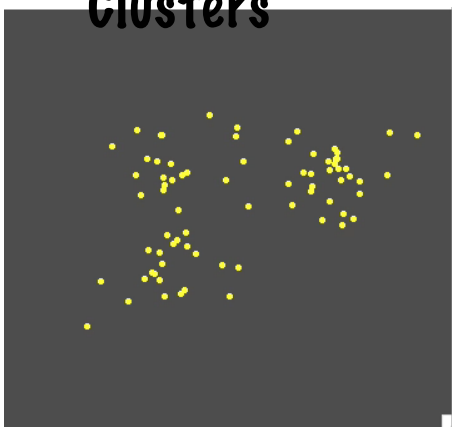


- Grand tour: The coefficients are randomly varied, giving a random walk over the space of all projections.



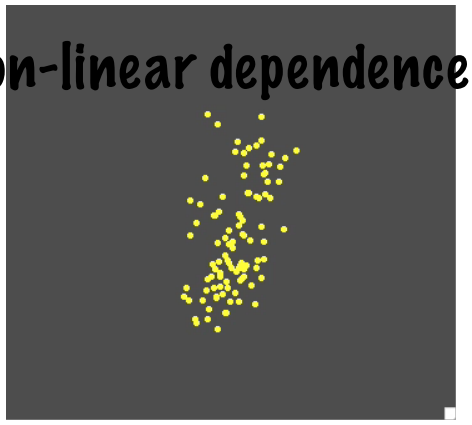
- Guided tour: The coefficients are chosen to maximize some criterion of interestingness, eg clusters, outliers, separation of classes.
- Manual tour: The user controls the coefficients of one of the variables. Good for exploring the importance of a single variable on the structure visible in the plot.

Clusters

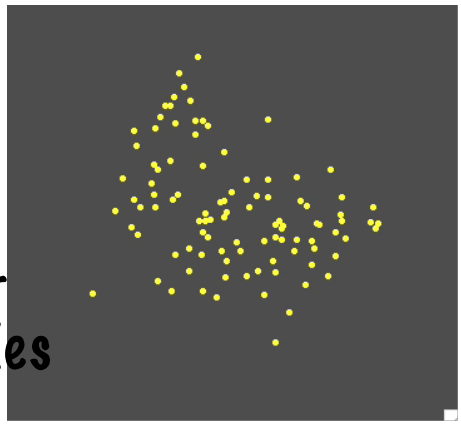


Tours

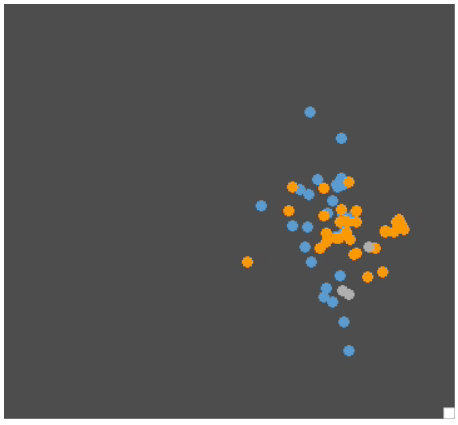
Non-linear dependence



Non-linear dependence + noise variables



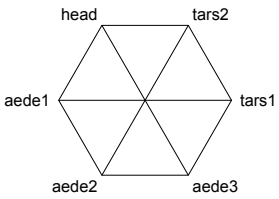
Class differences, outliers, non-linear dependence

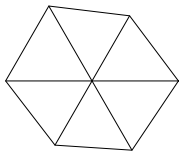


Tours

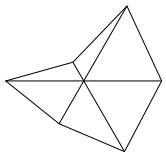
Something hidden

Icons



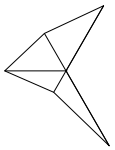


1 Concinna

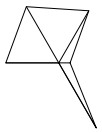


2 Concinna

- Each case is plotted separately.
- Each variable is mapped a feature of the icon



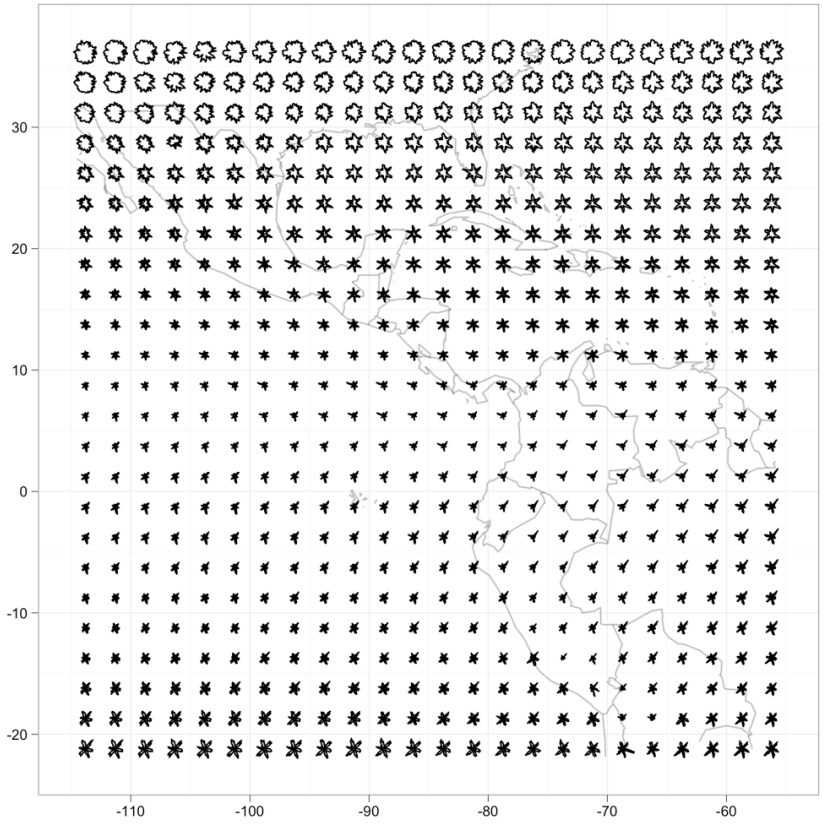
32 Heptapot.



33 Heptapot.

67 Heikert.

68 Heikert.



Icons

Example: Monthly ozone for 6 years over central America.

What do you see?

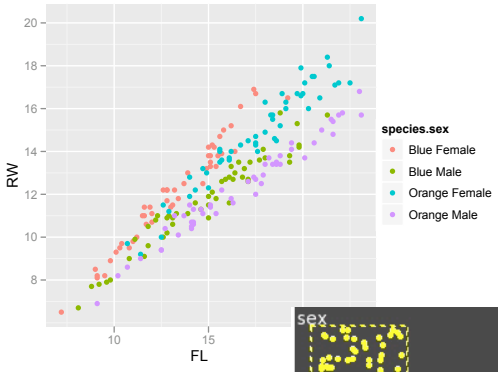
Multiple linked plots

Interactive graphics (direct manipulation graphics) allow the user to brush plot elements and convey this selection to other plots.

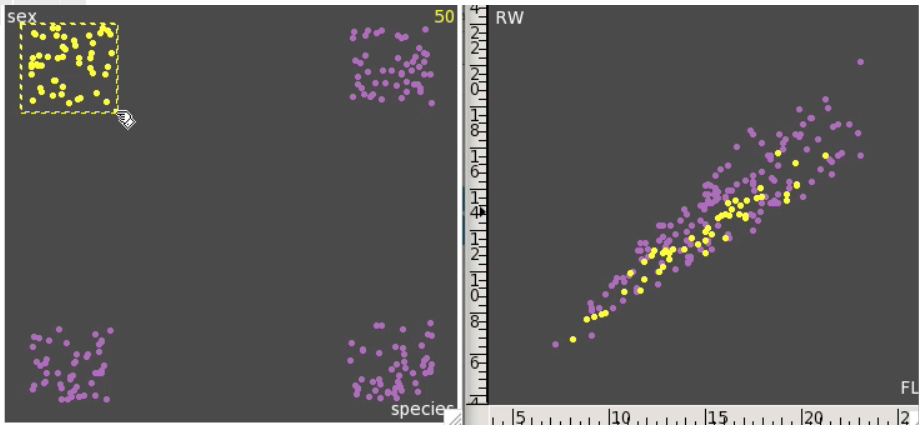
Terminology

Brushing (one-to-one)	
Brushing (categorical variable)	
Brushing (one point to one line)	
Transient brushing	
Persistent painting	
Identification	
Scaling	

This 1-1 Linking

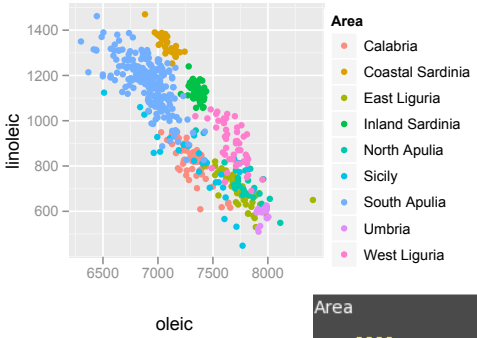


OR This?

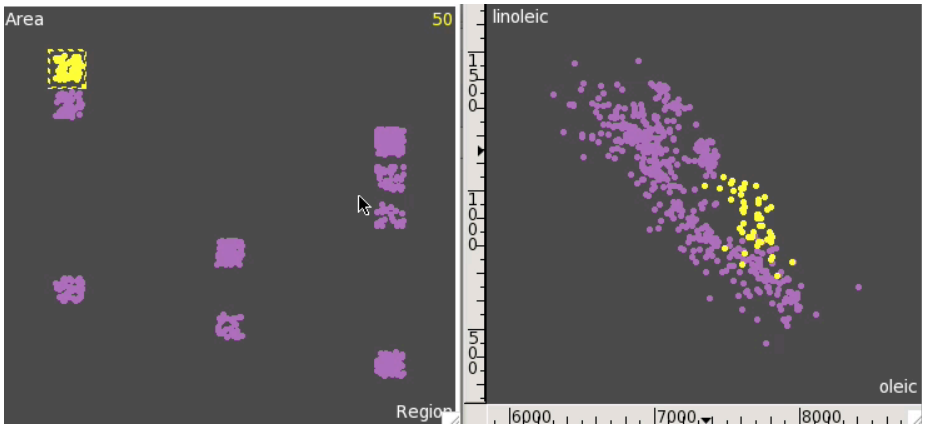


What do you learn differently?

This 1-1 Linking



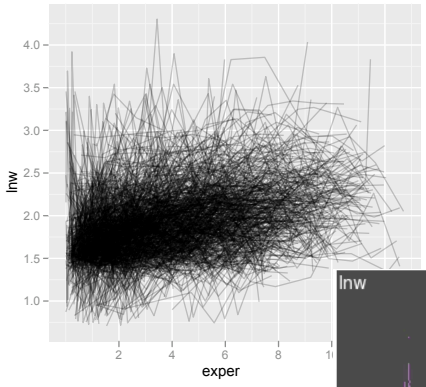
OR This?



What do you learn differently?

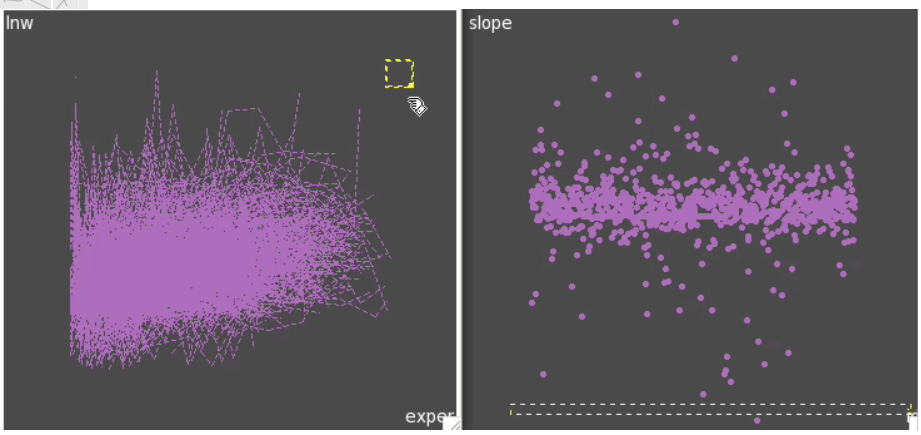
Categorical Variable Linking

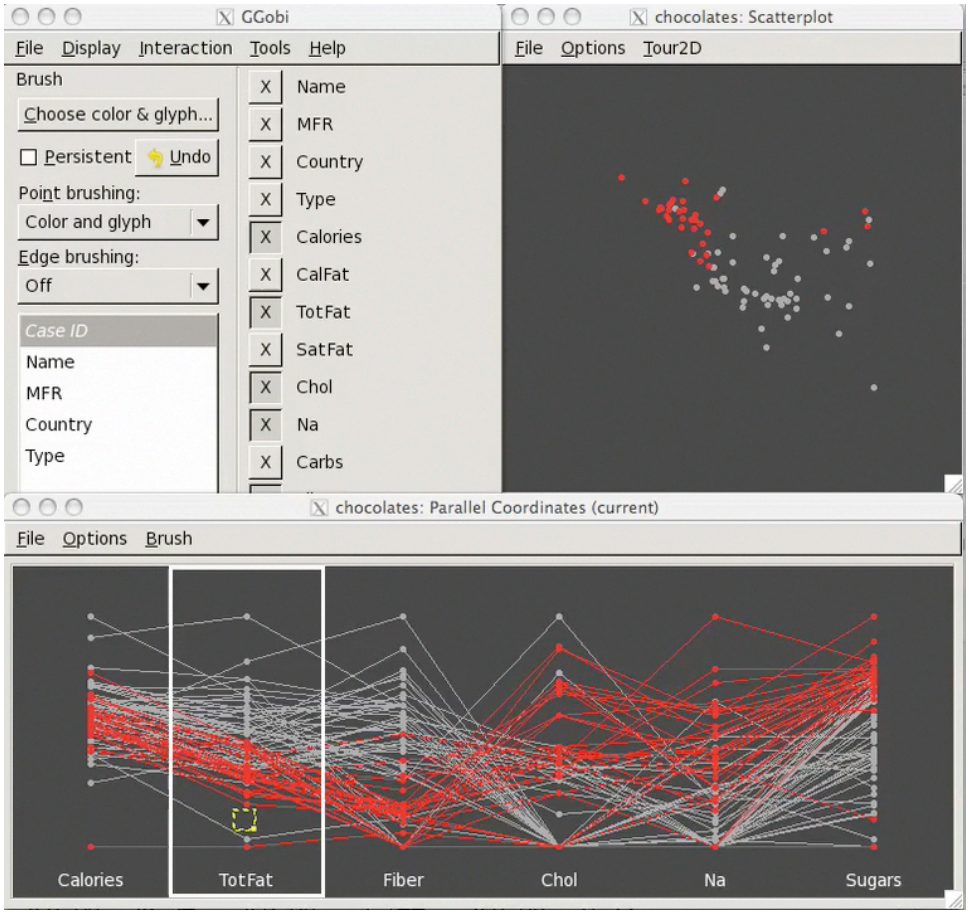
This



Longitudinal measurements on wages with workforce experience: 888 subjects.

OR This?

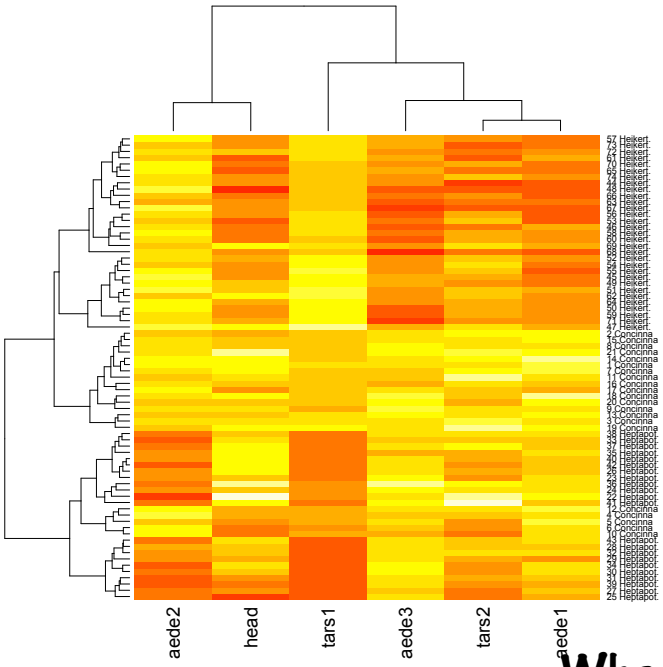




Example

Exploring dark chocolates that look more like milk chocolates.

One more DON't



Heatmaps: each cell of the data matrix is colored according to value. Rows and columns sorted to get most similar together.

How many clusters? Describe the clusters?

What's wrong with this plot?