

# Plotting multivariate data

---

Statistics 407  
ISU

# Multivariate data and graphics

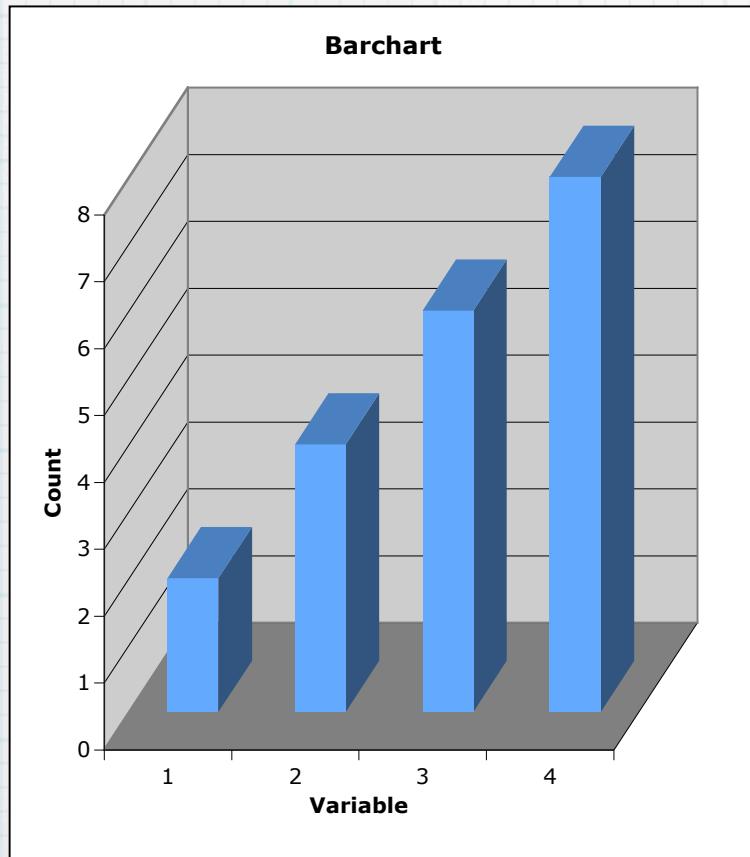
With multivariate data we want to understand the associations between multiple variables, which might be considered to be understanding the shape of the data in high-dimensional space.

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

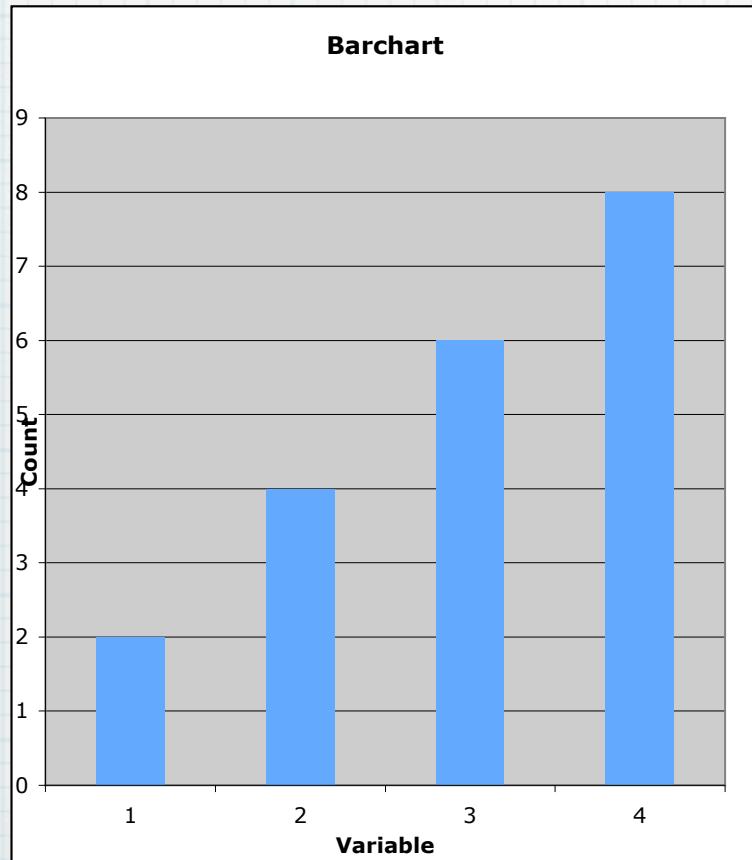
Eg Examining the number in each level of one of the label variables, prior to doing classification or MANOVA



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

# One variable - categorical

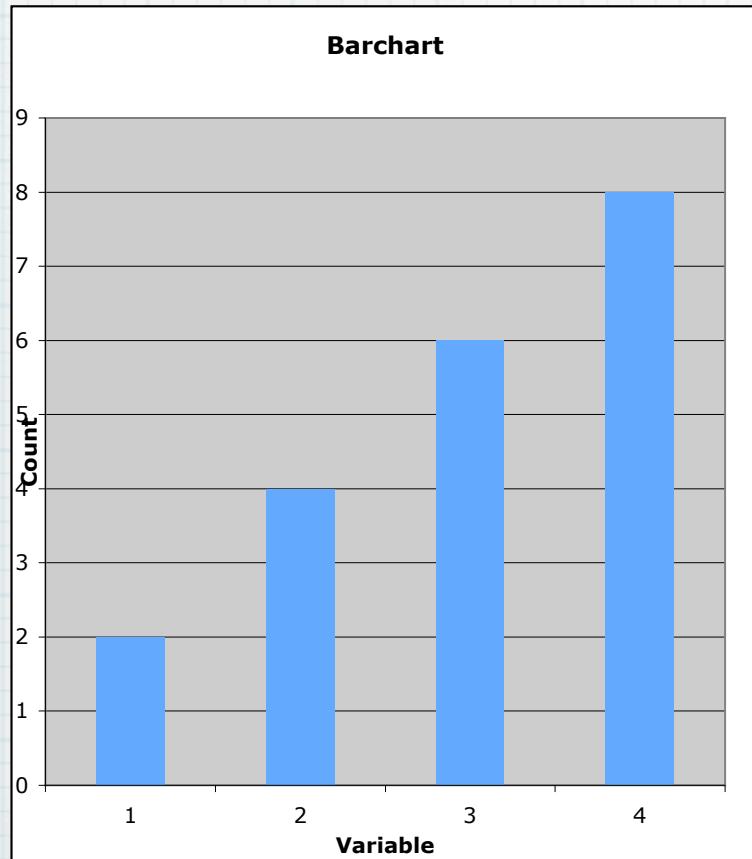
Eg Examining the number in each level of one of the label variables, prior to doing classification or MANOVA



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

# One variable - categorical

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



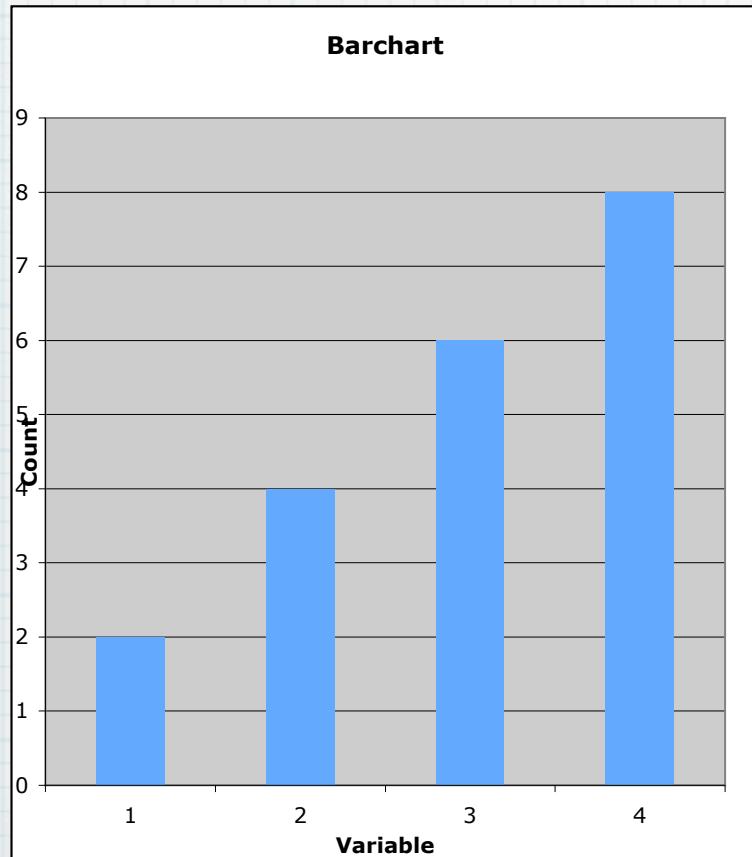
\* What do you think the counts are for each of the categories?

\* Naomi Robbins (2004)  
“Creating More Effective Graphs”  
Wiley

Counts are 2, 4, 6, 8

# 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

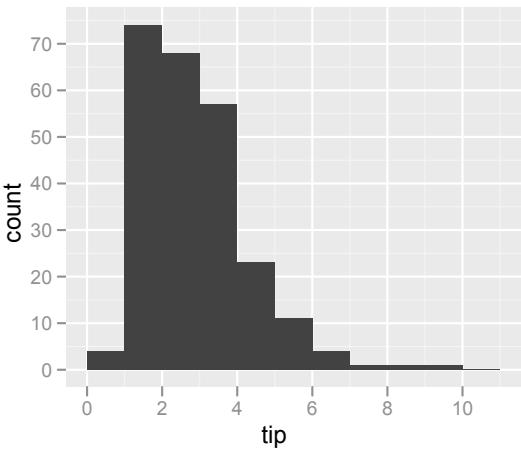
Counts are 2, 4, 6, 8

PLEASE DON'T USE 3D UNNECESSARILY

# One variable - real

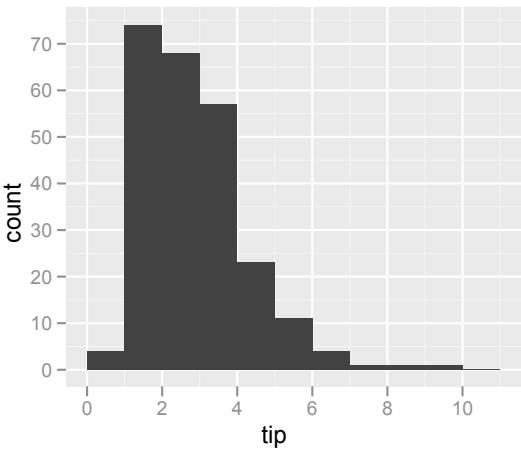
- \* Histogram: bin the values, and represent counts as bars. Use different bin sizes.
- \* Density: Like a smoothed histogram, estimate the density, represent as a curve.
- \* Dotplot: 1D scatterplot, values represented by dots.
- \* Boxplot: Representation of the 5 number summary, min/max, Q1/Q3, median.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

# One variable - real

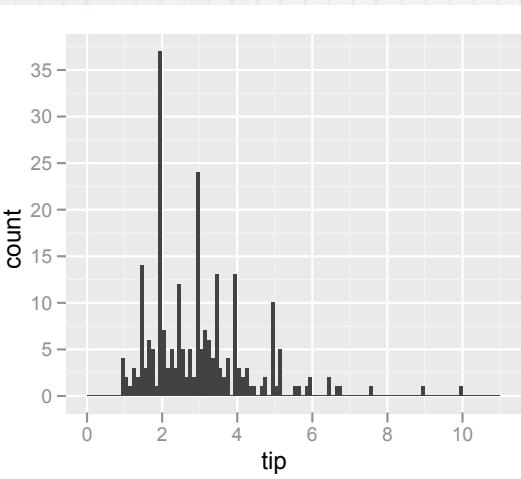


What do you learn  
about tips?

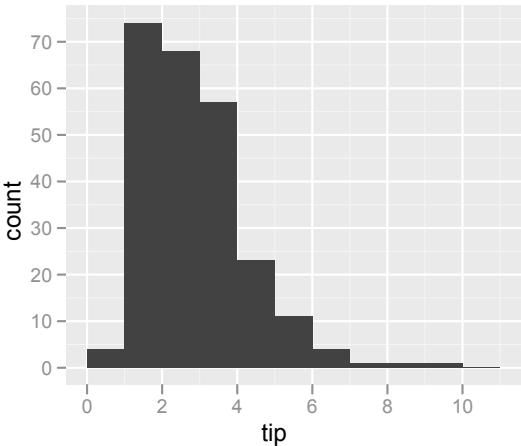
# One variable - real



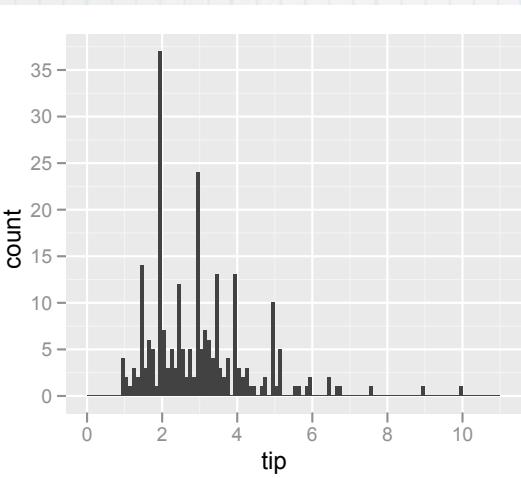
What do you learn  
about tips?



# One variable - real

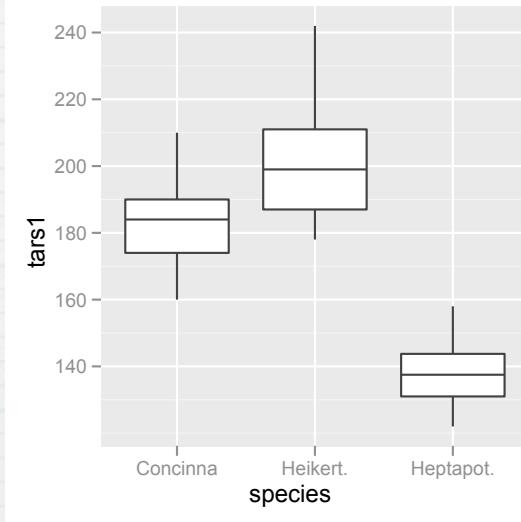


What do you learn  
about tips?

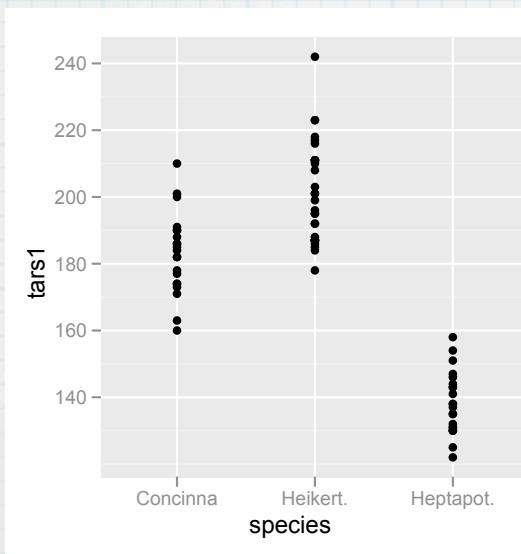


What else do you learn  
about tips?

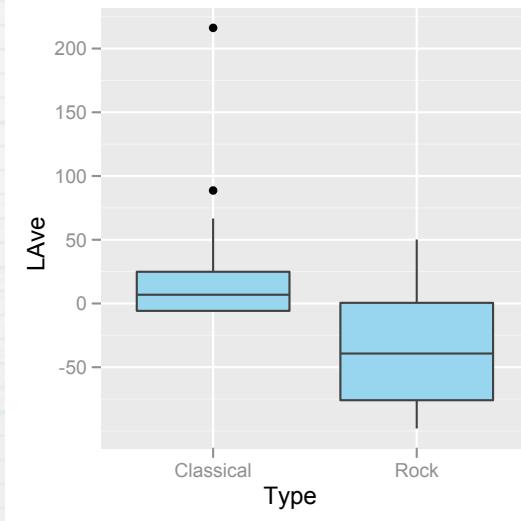
# One real + one categorical



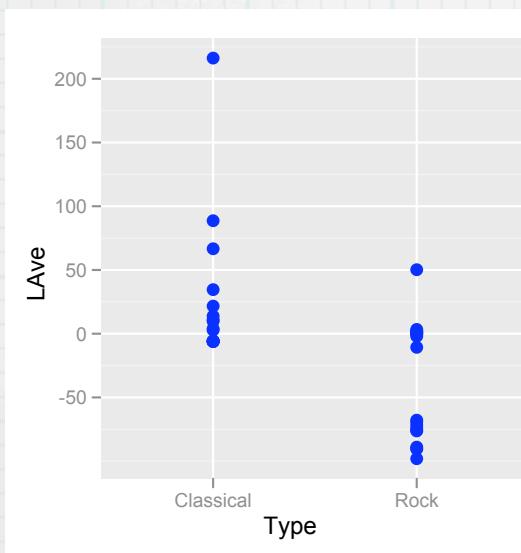
How do the species  
differ in the tars1  
values?



# One real + one categorical

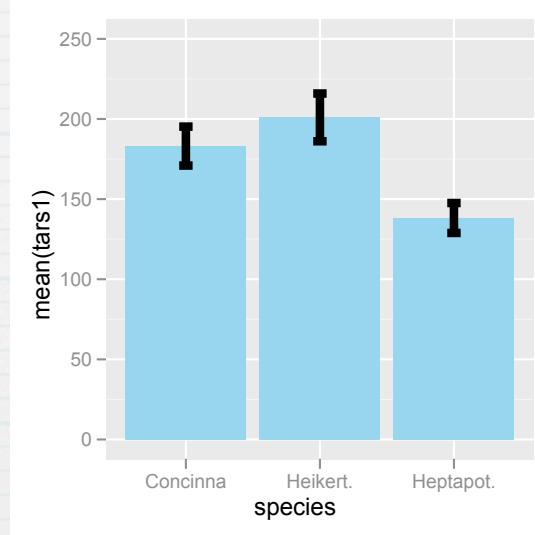


How does the type of music differ?

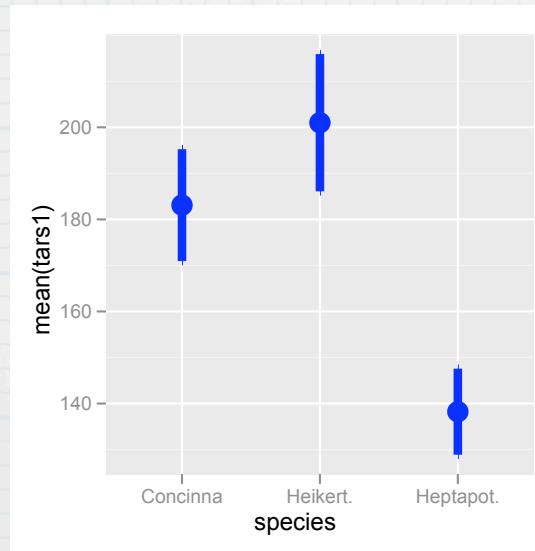


Do you learn anything different from the dotplot than the boxplot?

# Plotting Means

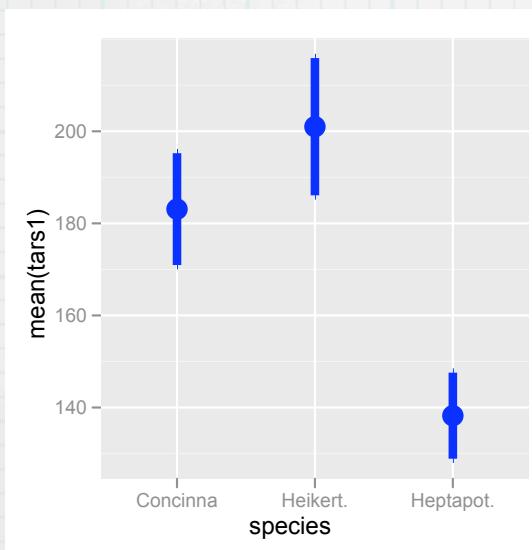
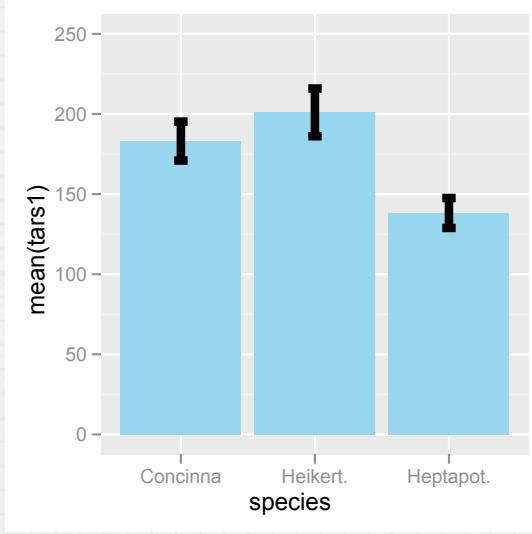


What's wrong with  
this plot?



What do we learn  
about differences  
between species?

# Plotting Means



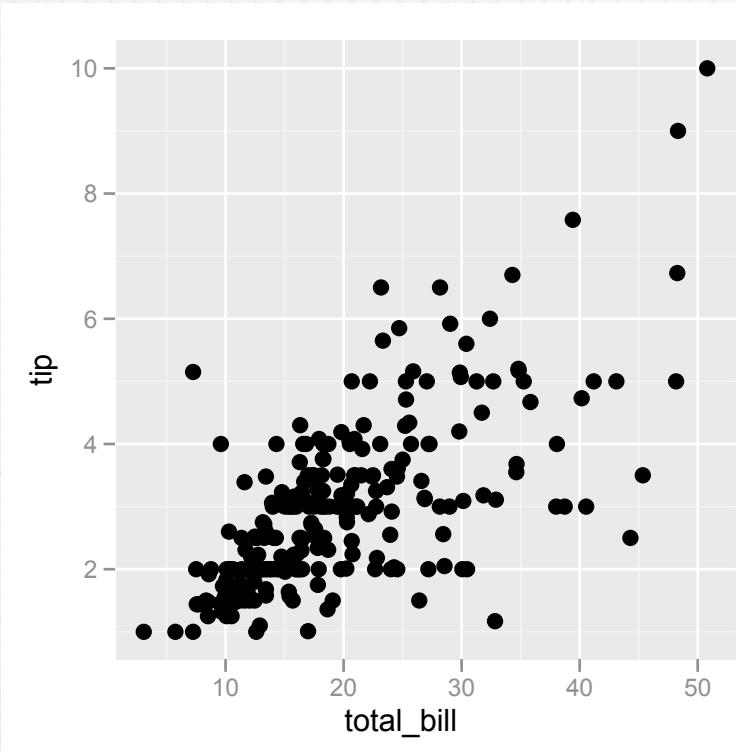
What's wrong with  
this plot?

Means are points  
estimates, use points to  
represent them.

What do we learn  
about differences  
between species?

# Two variables - real

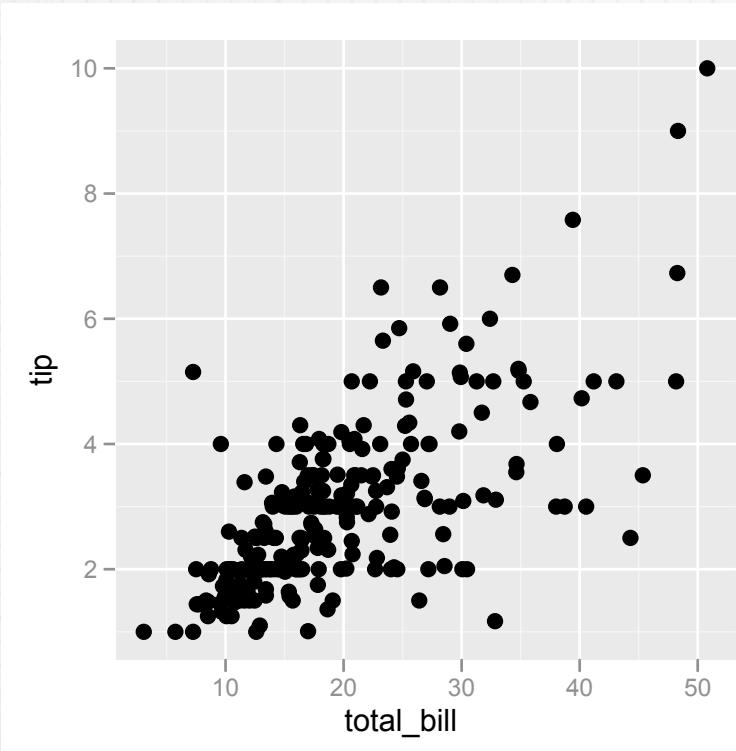
## \* Scatterplot



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

# Two variables - real

## \* Scatterplot

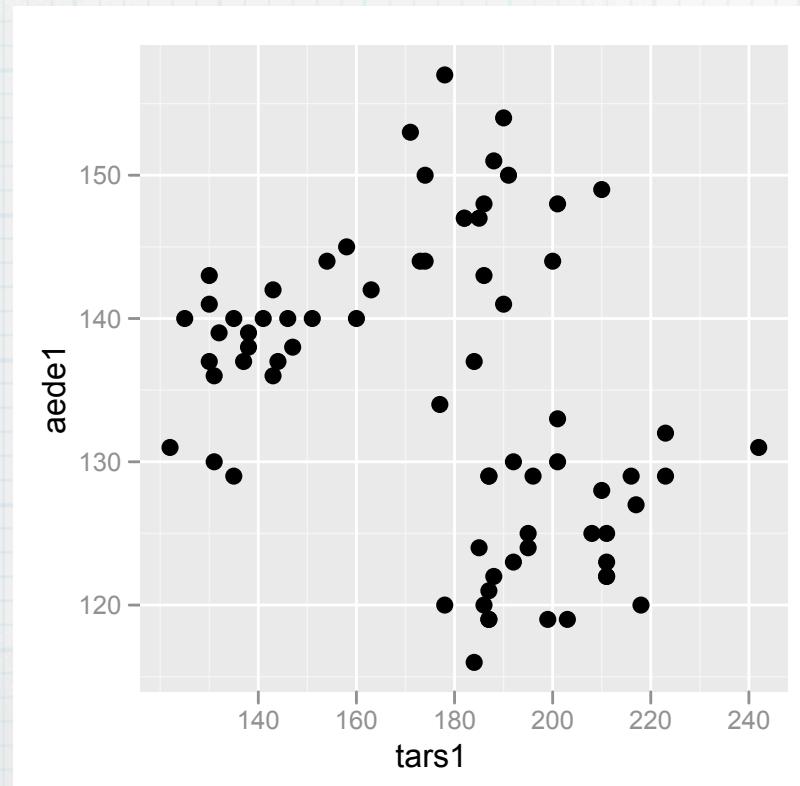


How is tip related  
to bill?  
How would you  
expect tip to be  
related to bill?

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

# Two variables - real

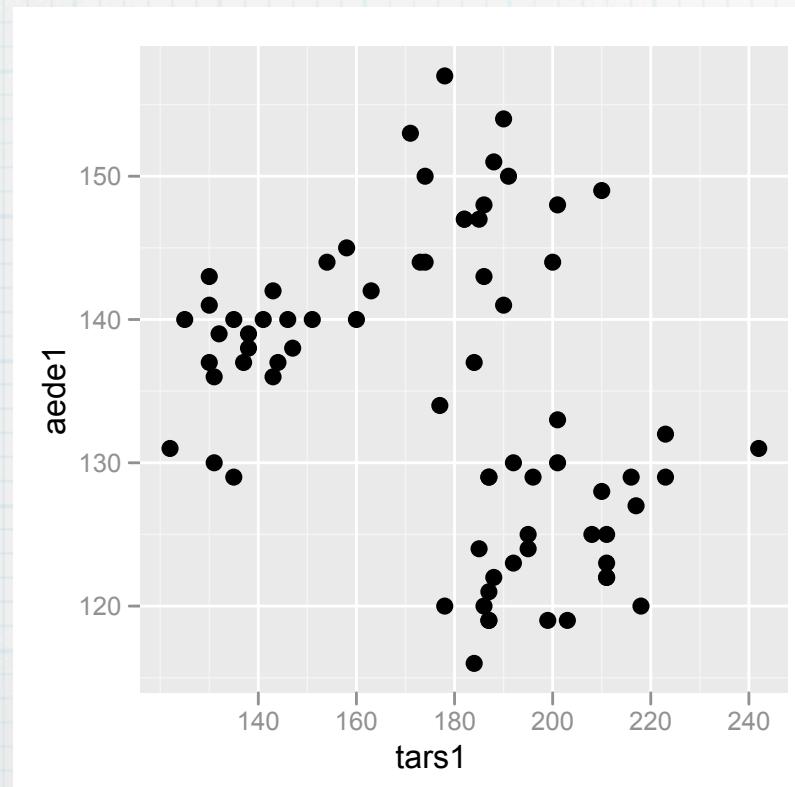
## \* Scatterplot



How is aede1  
related to tars1?

# Two variables - real

## \* Scatterplot

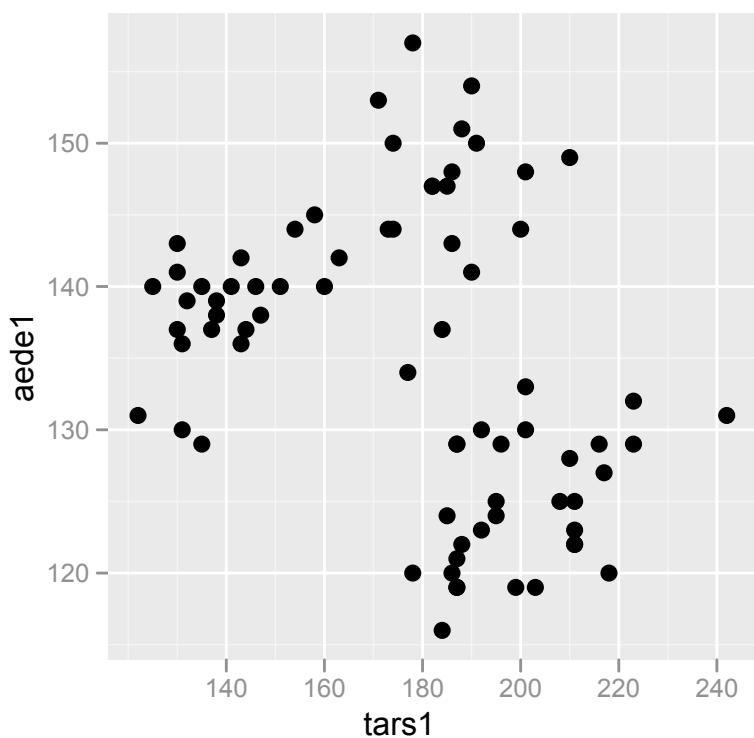


How is aede1  
related to tars1?

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

# Two variables - real

## \* Scatterplot



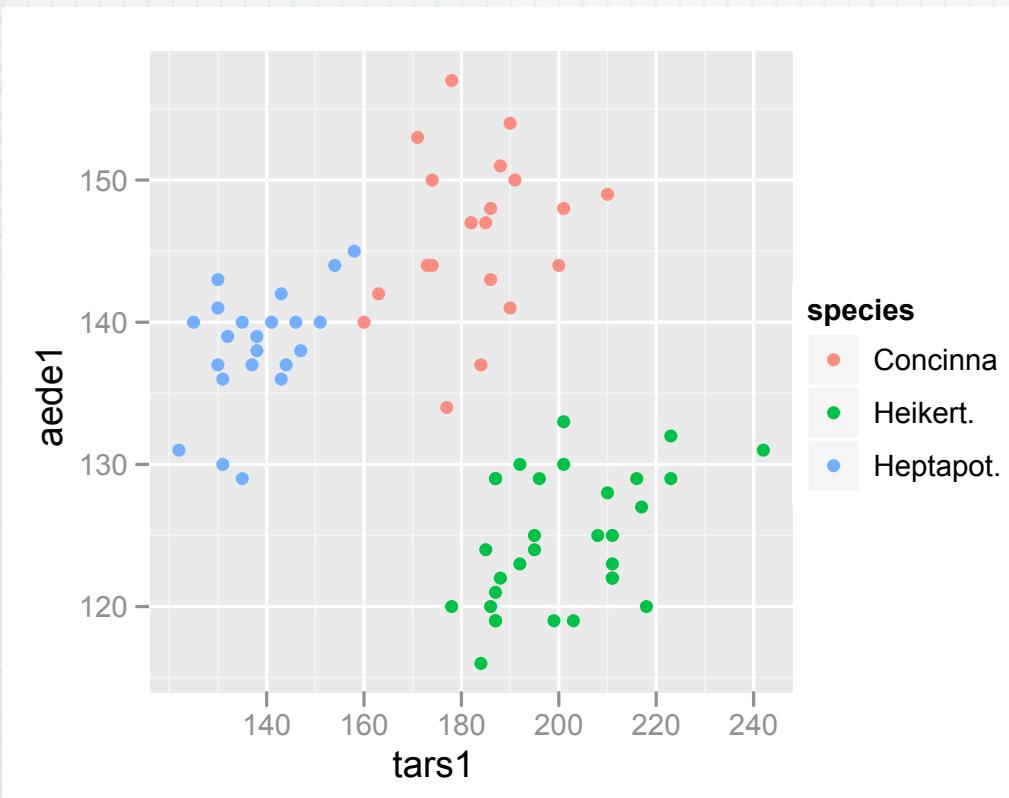
How is aede1  
related to tars1?

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

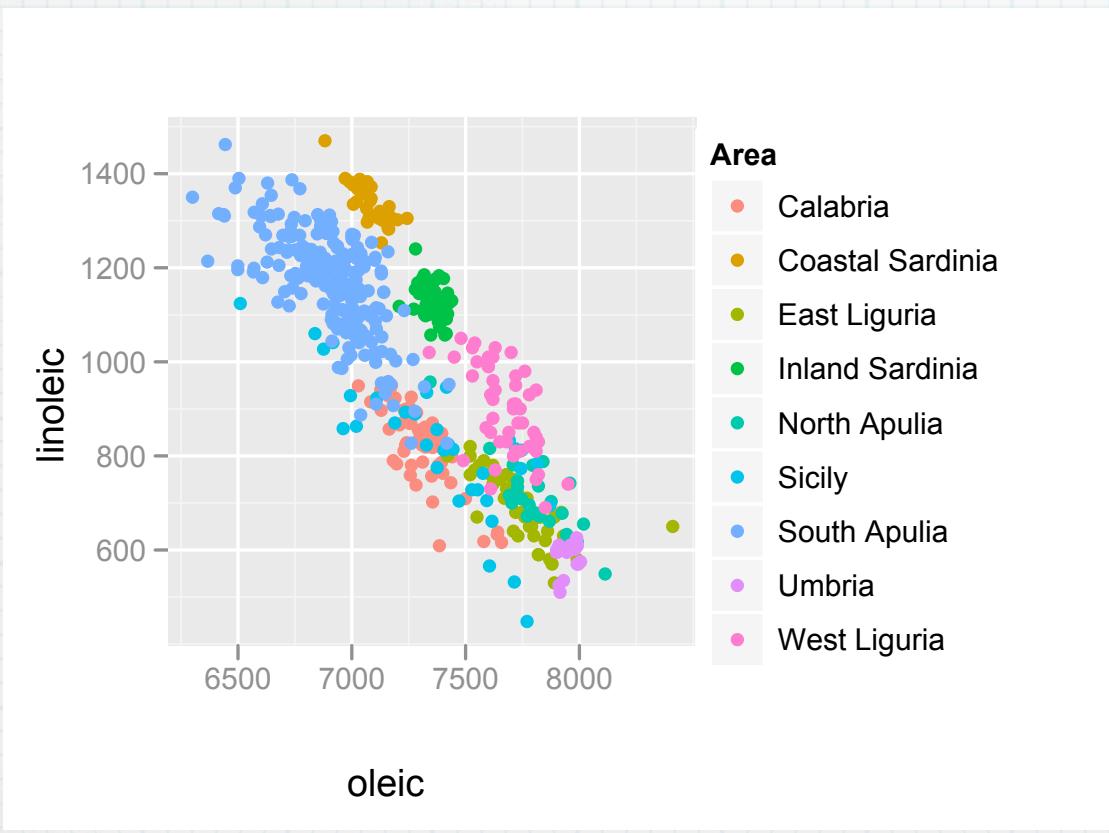
## \* Scatterplot



How is aede1  
related to tars1,  
by species?

# Two real + one categorical

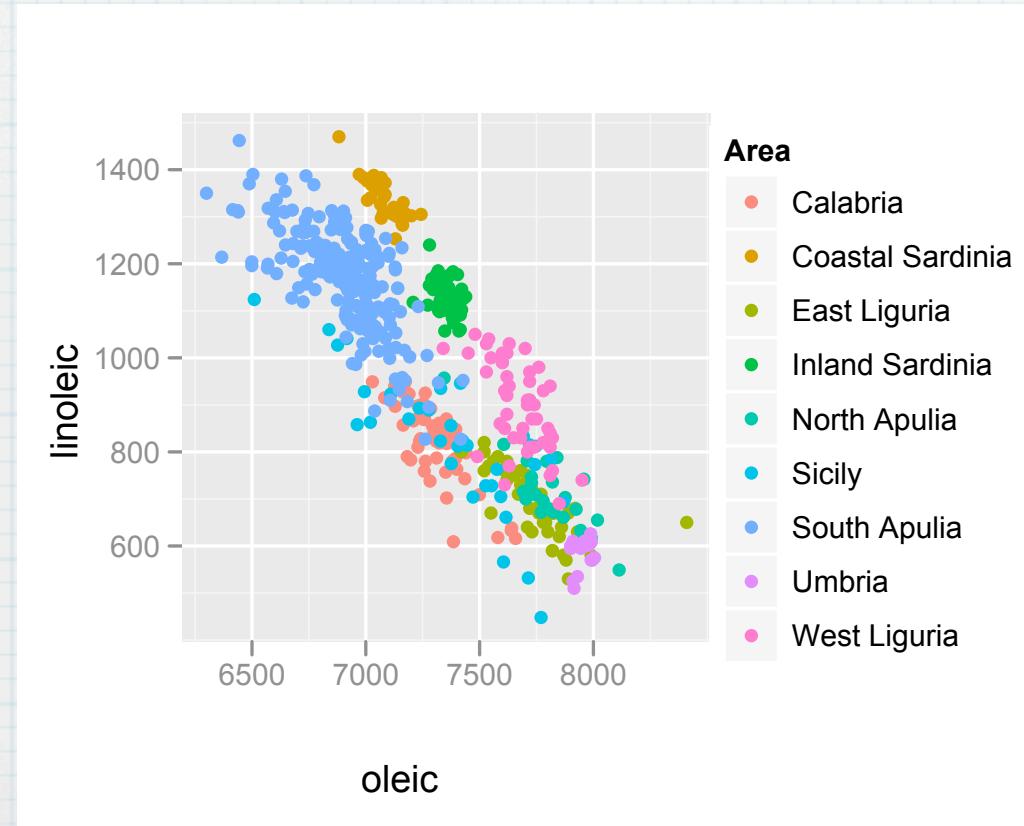
## \* Scatterplot



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

# Two real + one categorical

## \* Scatterplot

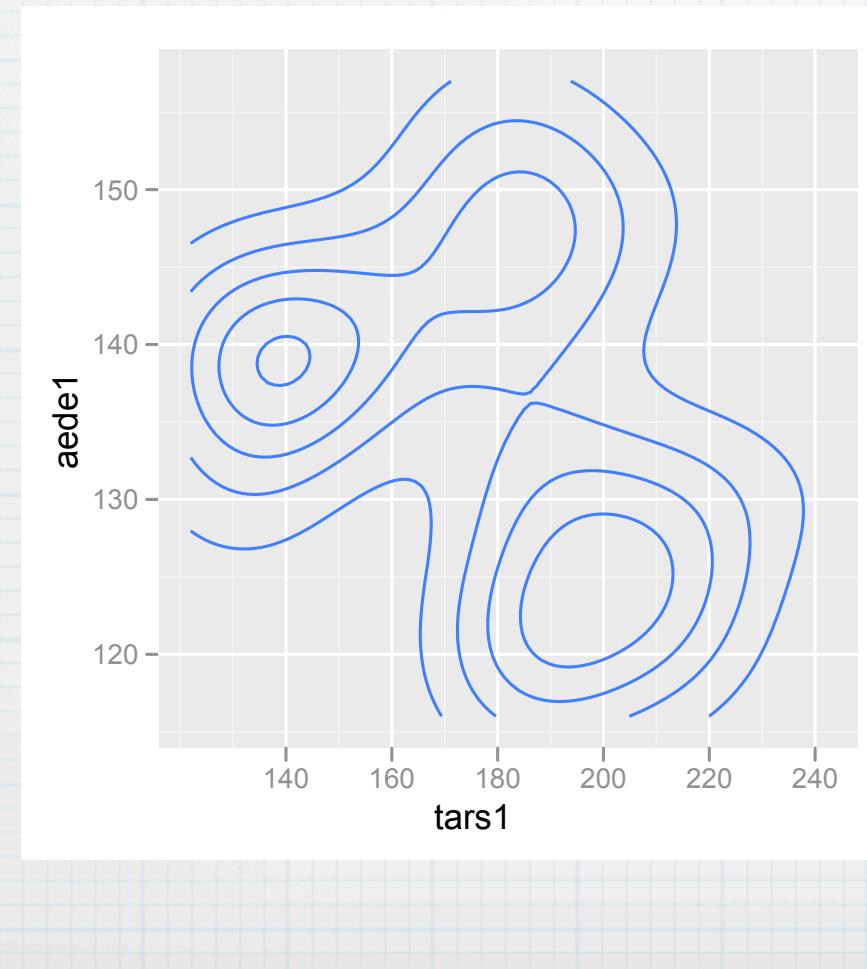


How is linoleic acid related to oleic acid, by area?

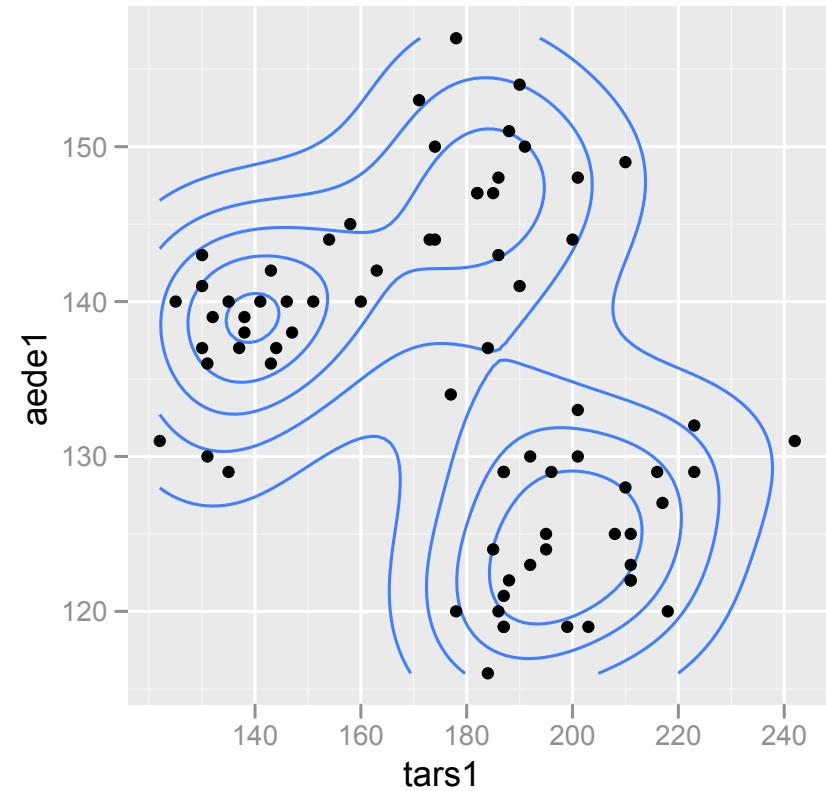
How many colors is too many?  
Generally only use 3-4 colors in a plot, only 3-4 categories.

# Two variables - real

\* 2D density

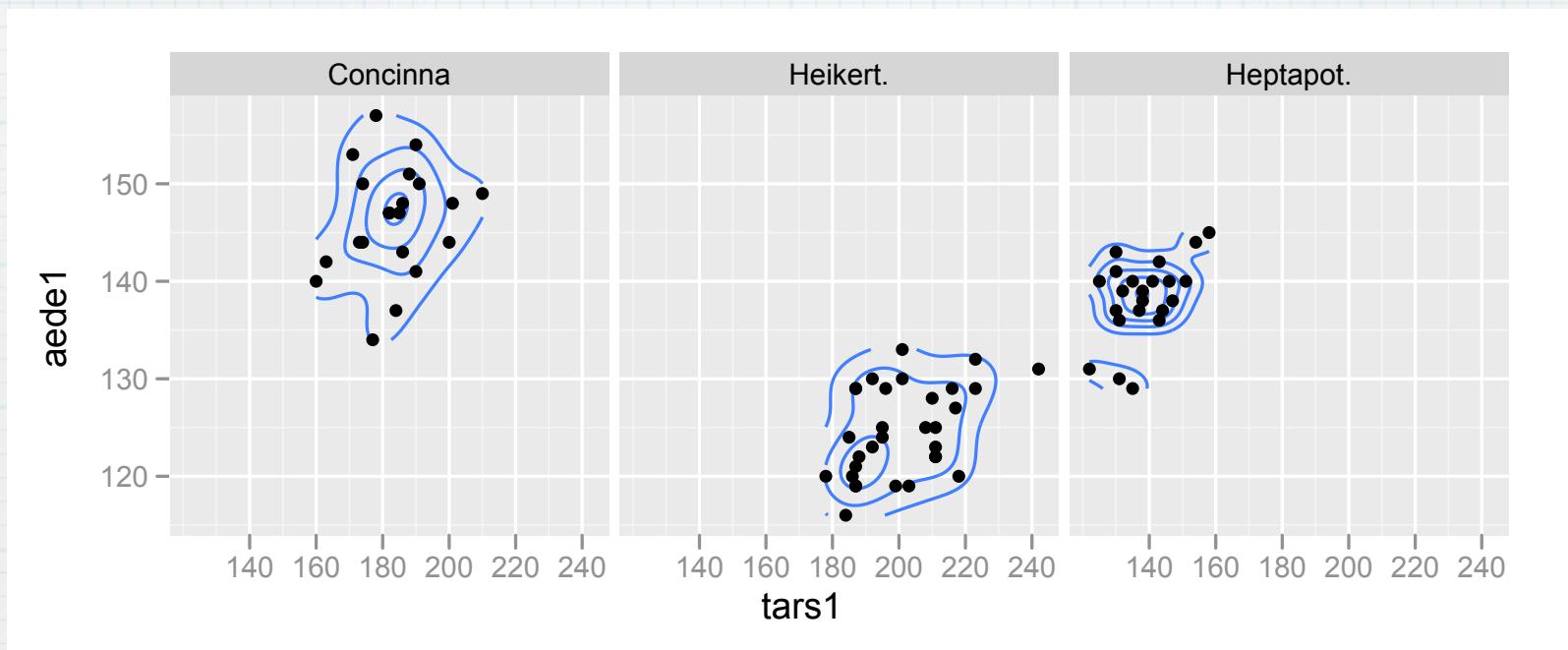


How is `aede1` related to `tars1`?



# Two real + one categorical

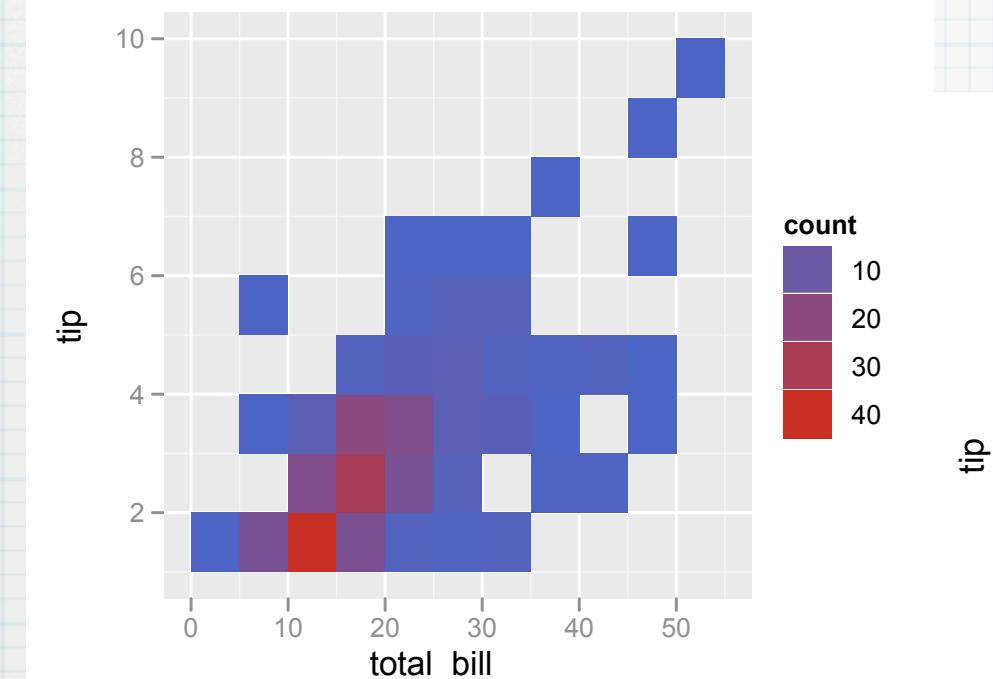
- \* Scatterplot + contour, faceted by species



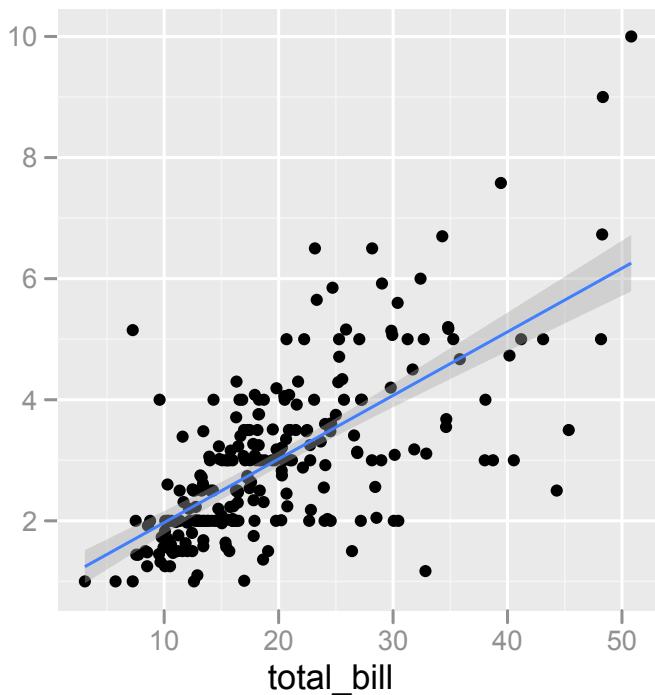
How is aede1 related to tars1, by species?

# Two variables - real

\* 2D histogram, scatterplot + linear model

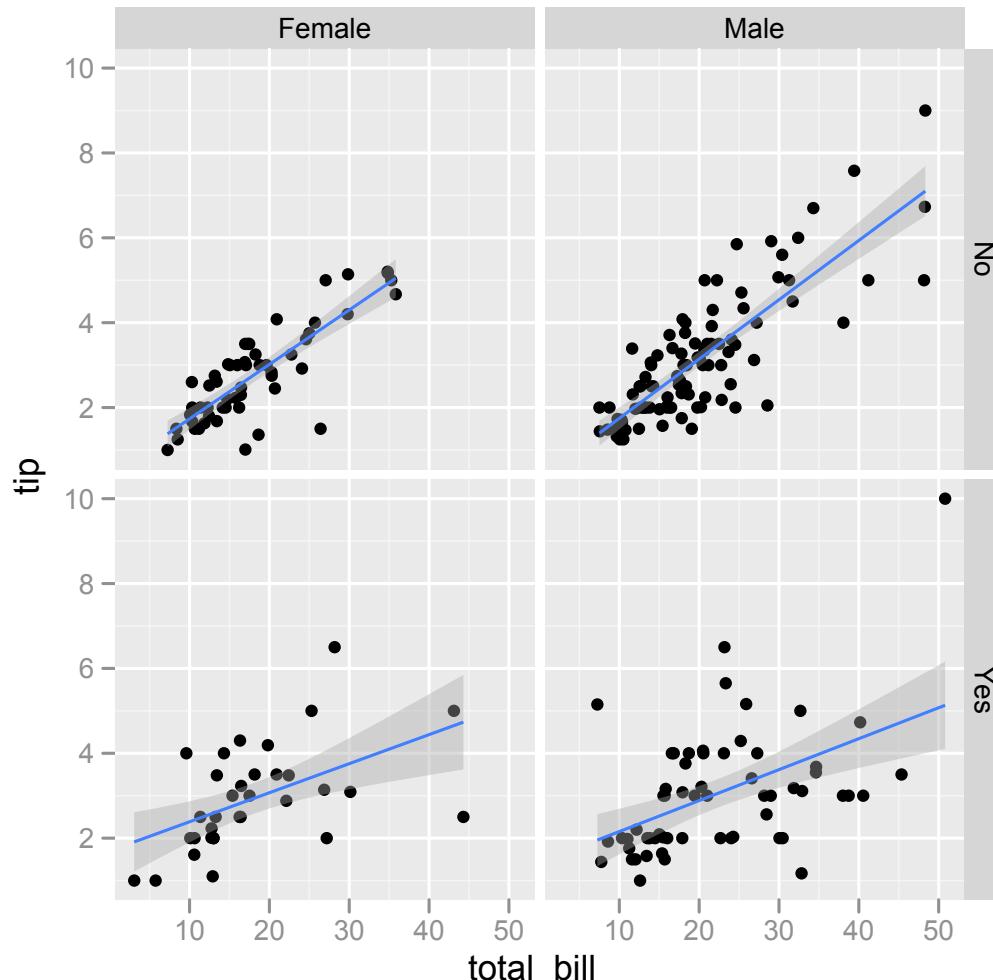


How is tip  
related to bill?



# Two real + one categorical

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



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

BEYOND 2D.....

# Scatterplot matrix

```
> 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

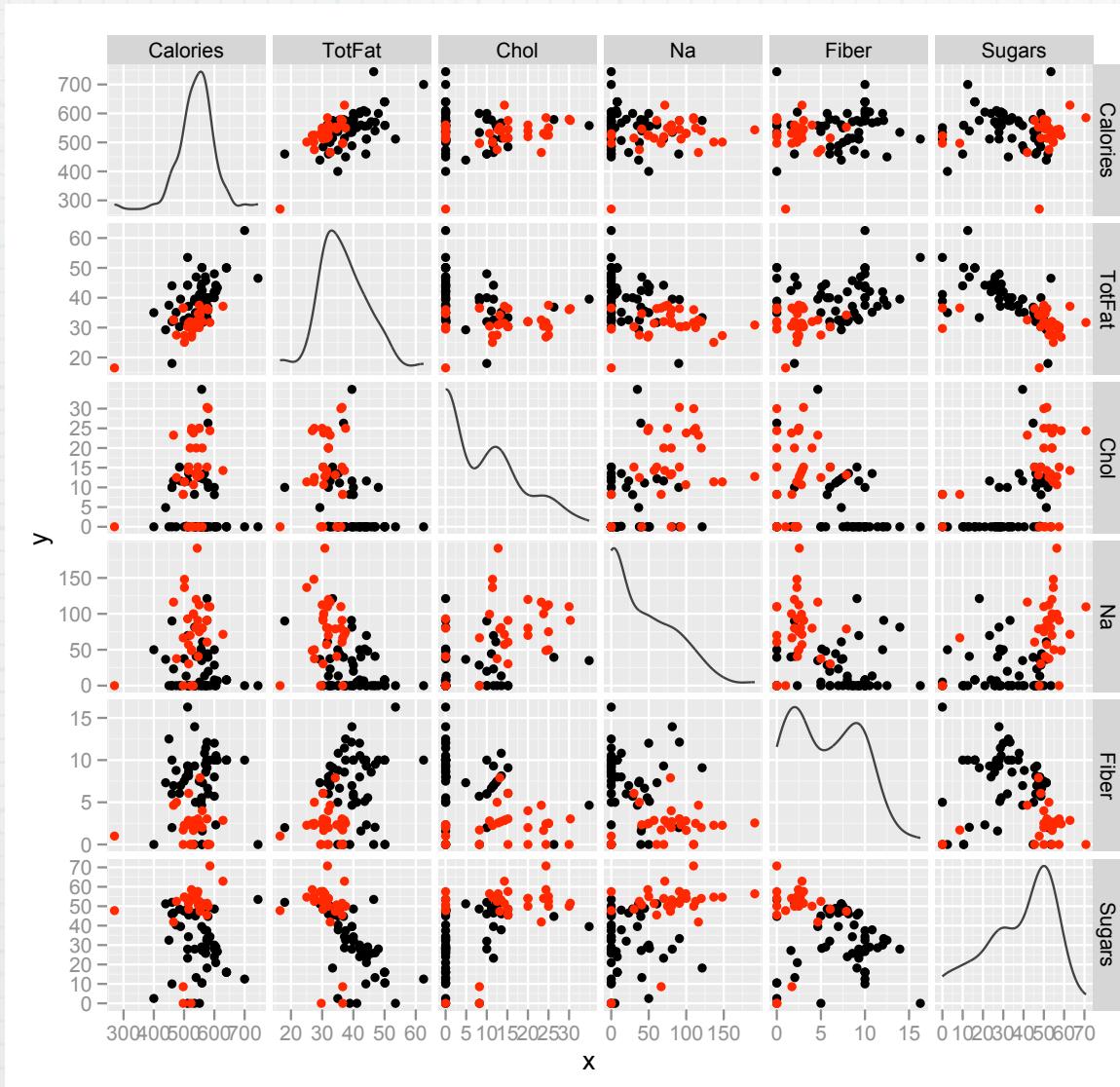
```
> cor(chocolates[chocolates$Type=="Dark",c(5,7,9,10,12,13)])
```

	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

Correlation matrix.

How do correlations differ by type of chocolate?

# Scatterplot matrix

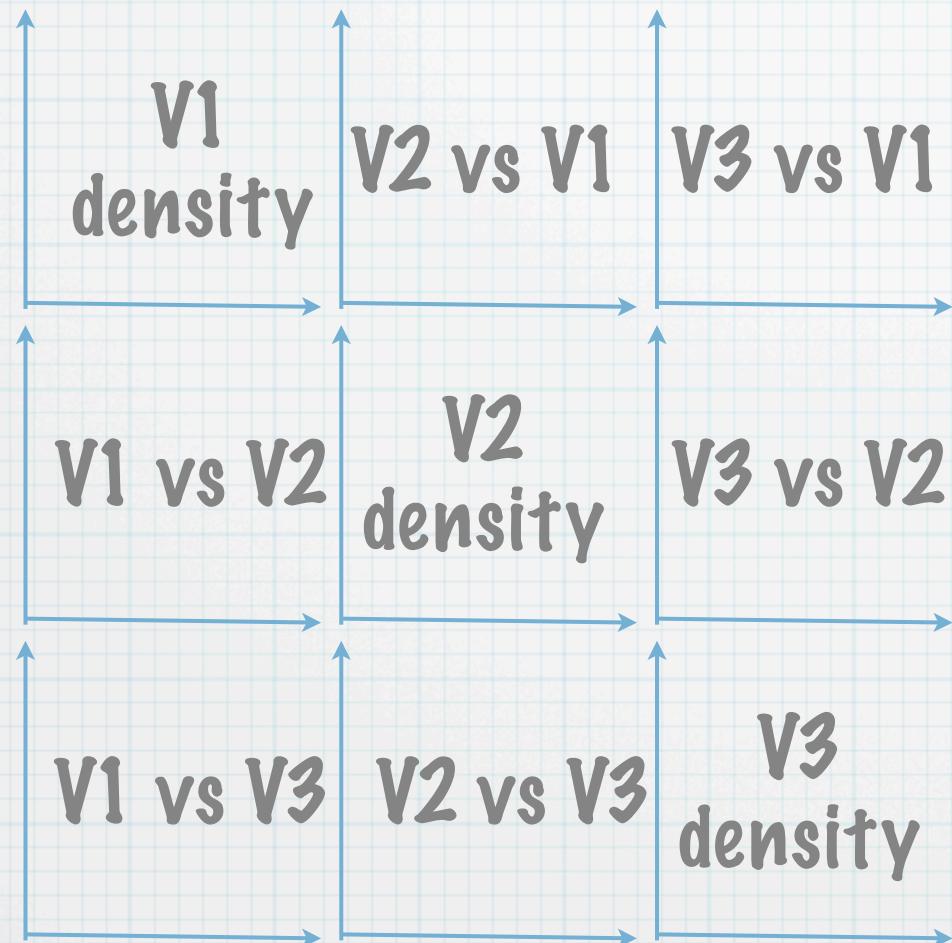


Red=milk  
Black=dark

All variables plotted pairwise, like correlation matrix.

What do you learn?

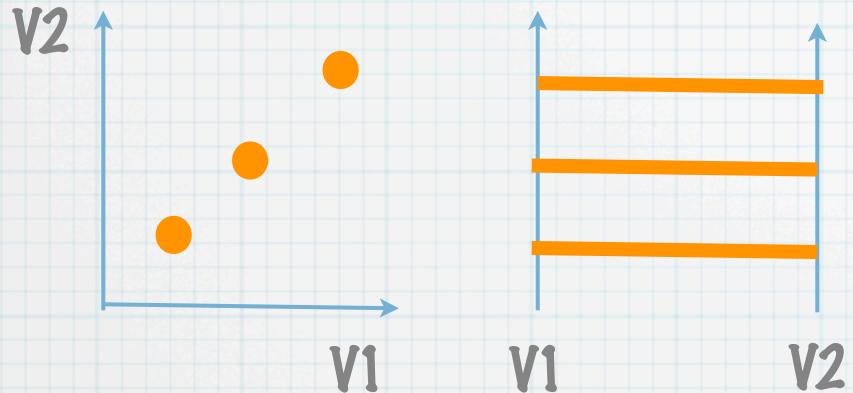
# Scatterplot matrix



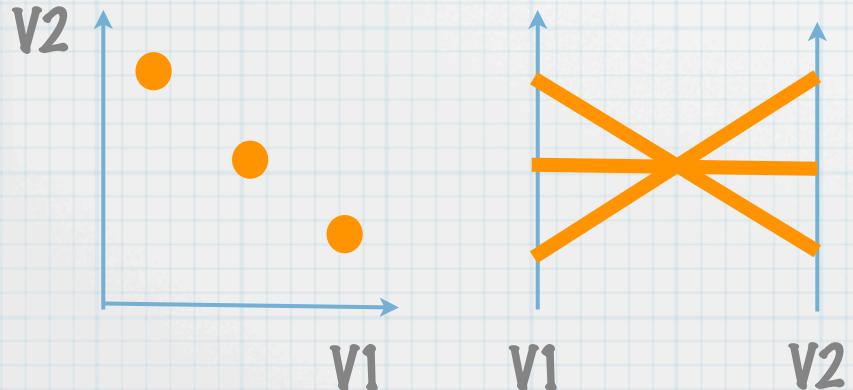
GENERAL  
Pairwise  
dependence only.

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

# Parallel coordinates



Orthogonal      Parallel



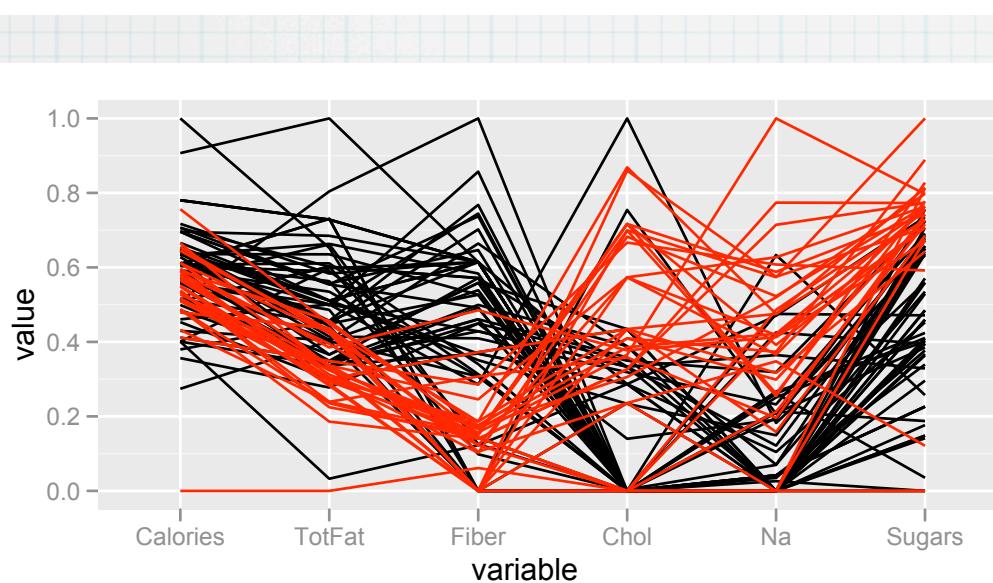
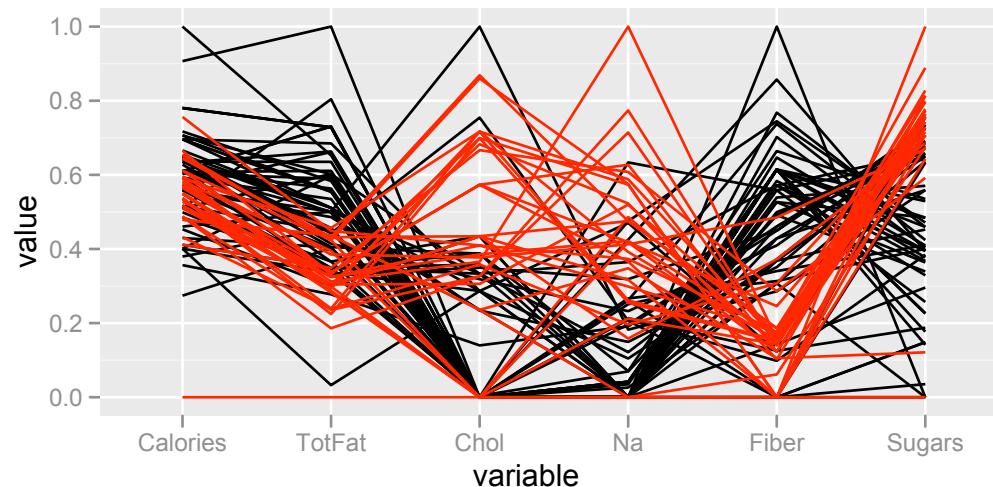
## GENERAL

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



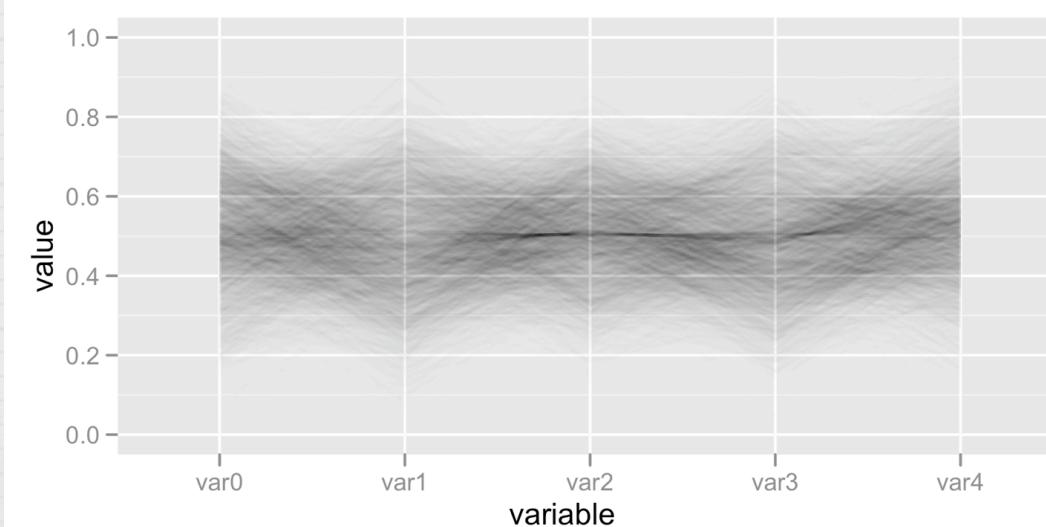
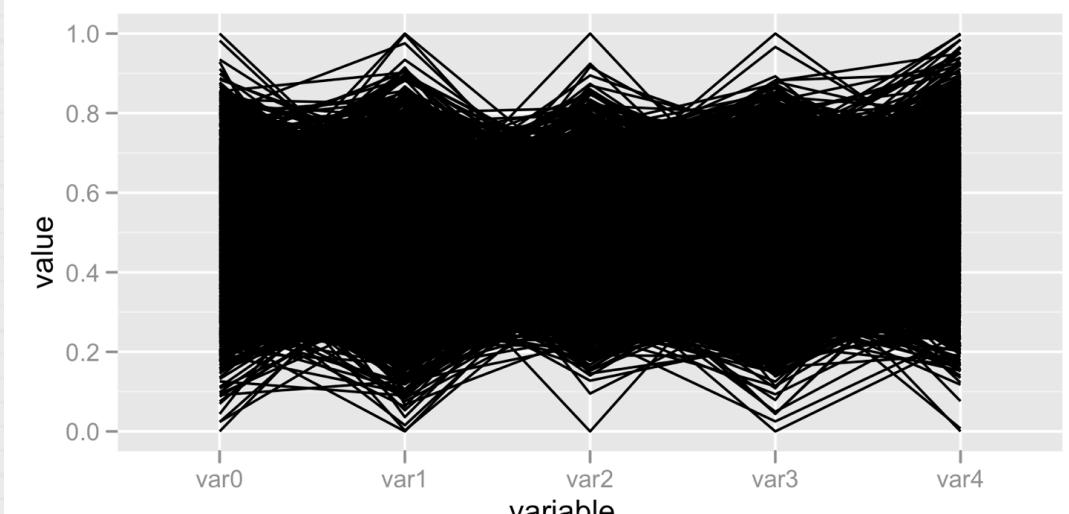
Red=milk  
Black=dark

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

What's the difference between milk and dark chocolate?

N=3367

# Parallel coordinates



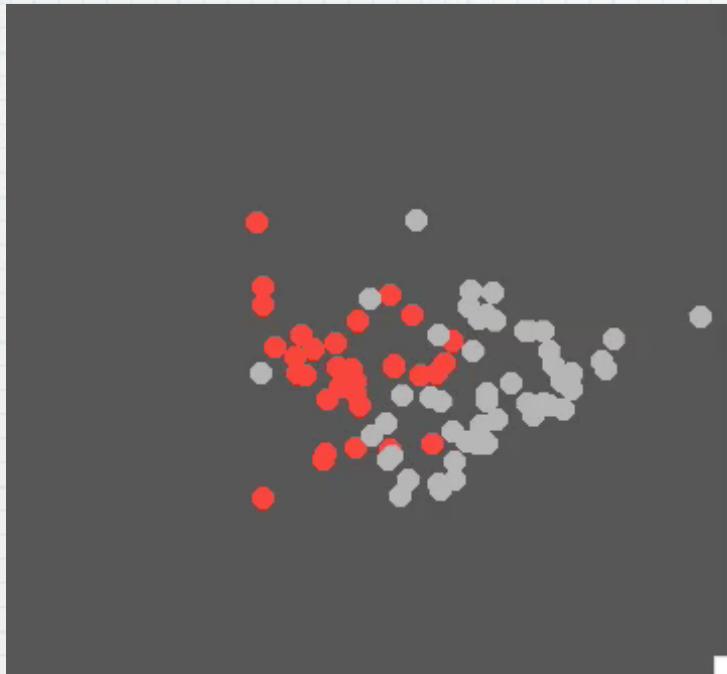
Disadvantage:

Large n means a lot  
of overplotting.

Using “alpha  
blending” might  
alleviate this.

What do you see  
now?

# 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

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

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  $\mathbf{X}_{n \times p}$

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

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.

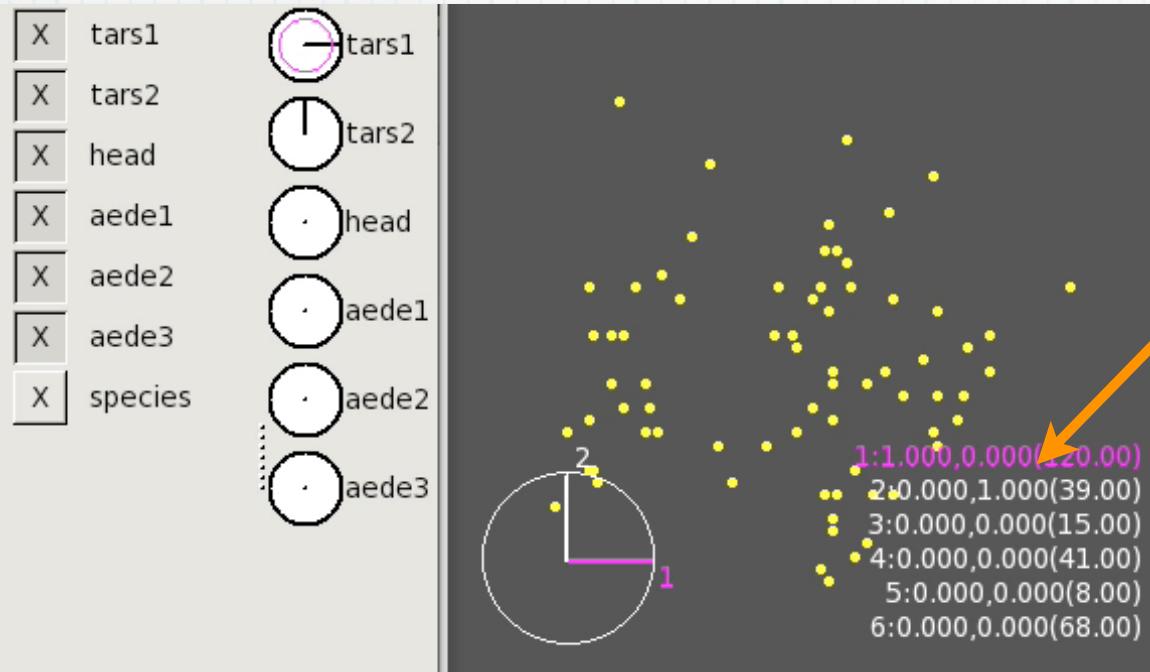
# Visual representation of coefficients (axes)

# Tours



Projection  
coefficients and  
scaling

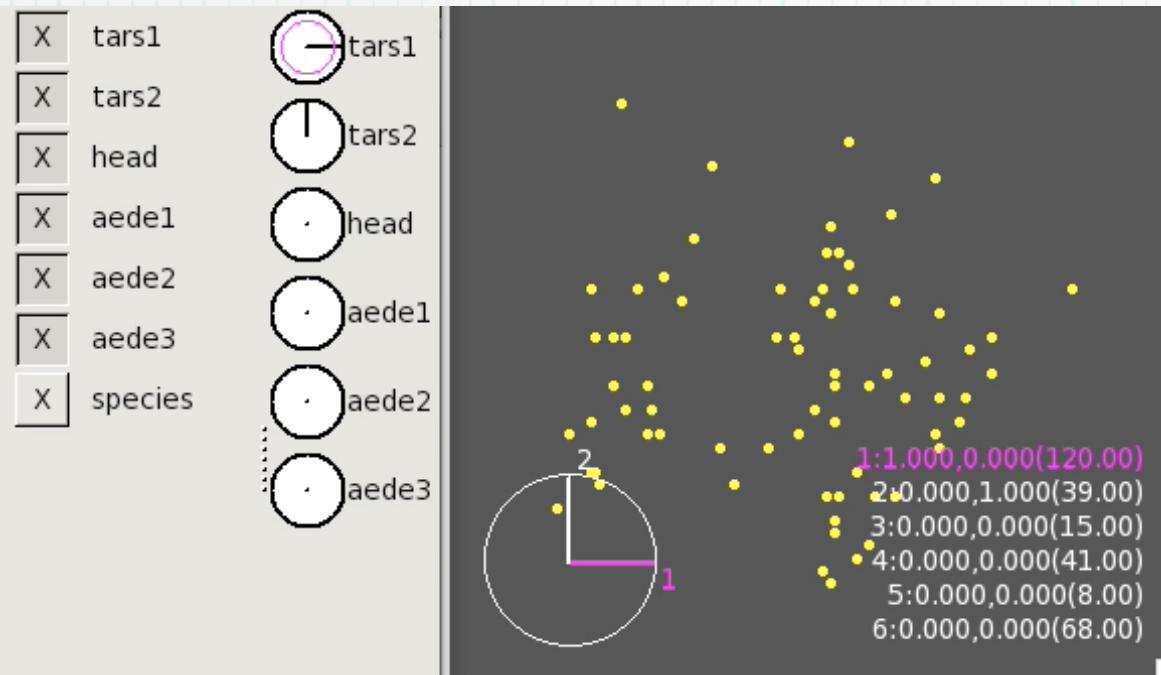
# Tours



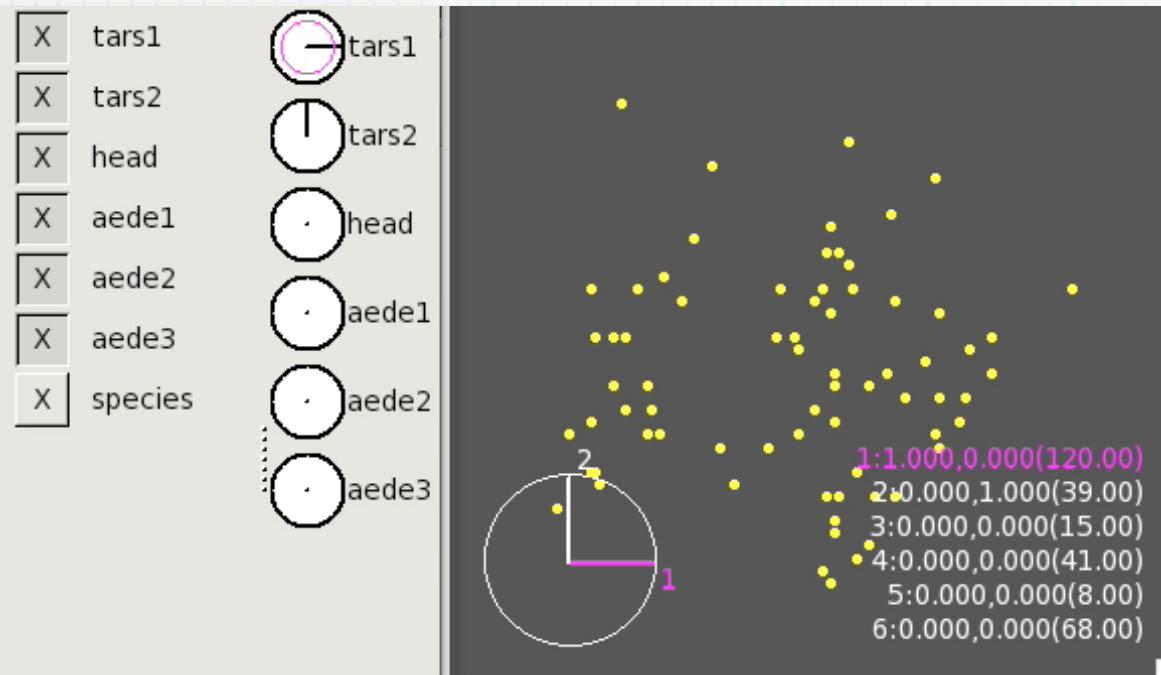
Projection  
coefficients and  
scaling

1:1.000,0.000(120.00)  
2:0.000,1.000(39.00)  
3:0.000,0.000(15.00)  
4:0.000,0.000(41.00)  
5:0.000,0.000(8.00)  
6:0.000,0.000(68.00)

# Tours



# Tours



$$\mathbf{A} = \begin{bmatrix} 0.548 & -0.152 \\ -0.152 & 0.754 \\ 0.185 & 0.363 \\ 0.031 & 0.463 \\ 0.793 & 0.116 \\ -0.110 & -0.222 \end{bmatrix}$$

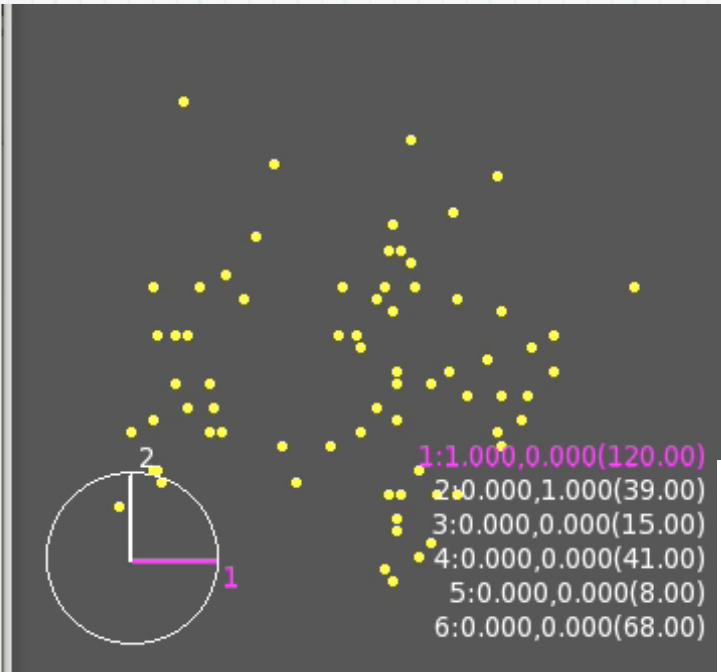
scaling for each variable

$$\begin{bmatrix} 0.548/120 & -0.152/120 \\ -0.152/39 & 0.754/39 \\ 0.185/15 & 0.363/15 \\ 0.031/41 & 0.463/41 \\ 0.793/8 & 0.116/8 \\ -0.110/68 & -0.222/68 \end{bmatrix}$$

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

# Tours

X	tars1	
X	tars2	
X	head	
X	aede1	
X	aede2	
X	aede3	
X	species	

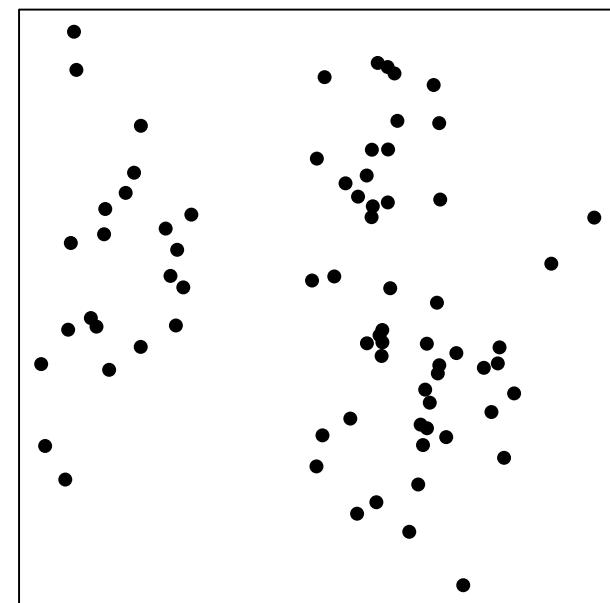


$$A = \begin{bmatrix} 0.548 & -0.152 \\ -0.152 & 0.754 \\ 0.185 & 0.363 \\ 0.031 & 0.463 \\ 0.793 & 0.116 \\ -0.110 & -0.222 \end{bmatrix}$$

scaling for each variable

$$\begin{bmatrix} 0.548/120 & -0.152/120 \\ -0.152/39 & 0.754/39 \\ 0.185/15 & 0.363/15 \\ 0.031/41 & 0.463/41 \\ 0.793/8 & 0.116/8 \\ -0.110/68 & -0.222/68 \end{bmatrix}$$

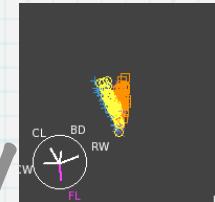
Multiply the data matrix by this to get the particular view of the data.



# Types of Tours

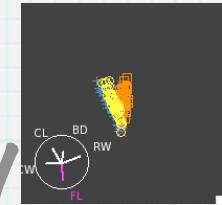
- \* **Grand tour:** The coefficients are randomly varied, giving a random walk over the space of all projections.
- \* **Guided tour:** The coefficients are chosen to maximize some criterion of interestingness, eg clusters, outliers, separation of classes.
- \* **Manual tour:** The user controls the coefficients of one of the variables. Good for exploring the importance of a single variable on the structure visible in the plot.

# Types of Tours

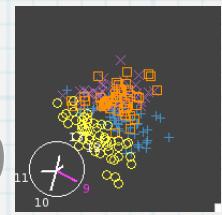


- \* **Grand tour:** The coefficients are randomly varied, giving a random walk over the space of all projections.
- \* **Guided tour:** The coefficients are chosen to maximize some criterion of interestingness, eg clusters, outliers, separation of classes.
- \* **Manual tour:** The user controls the coefficients of one of the variables. Good for exploring the importance of a single variable on the structure visible in the plot.

# Types of Tours

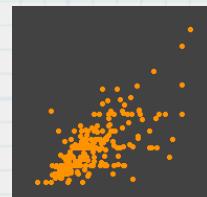
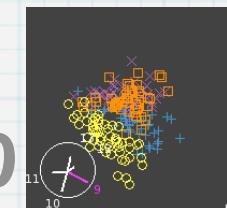
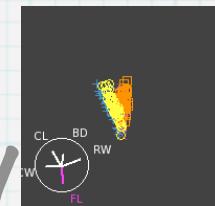


- \* **Grand tour:** The coefficients are randomly varied, giving a random walk over the space of all projections.
- \* **Guided tour:** The coefficients are chosen to maximize some criterion of interestingness, eg clusters, outliers, separation of classes.
- \* **Manual tour:** The user controls the coefficients of one of the variables. Good for exploring the importance of a single variable on the structure visible in the plot.

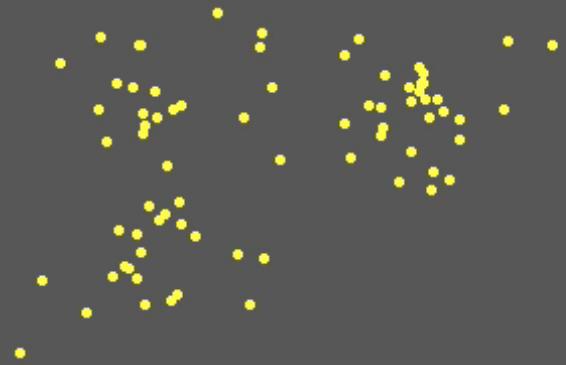


# Types of Tours

- \* **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

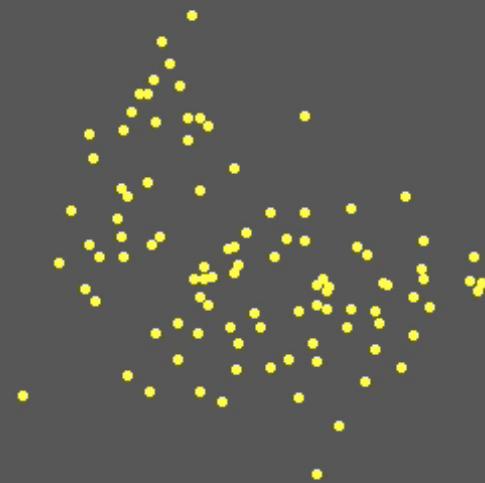


Non-linear dependence

Tours



Non-linear  
dependence +  
noise variables



Class differences,  
outliers, non-  
linear dependence



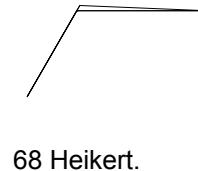
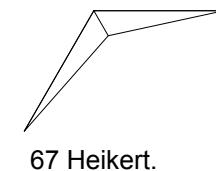
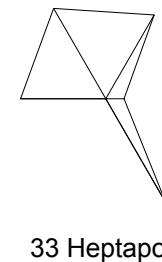
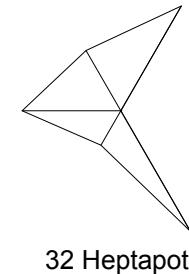
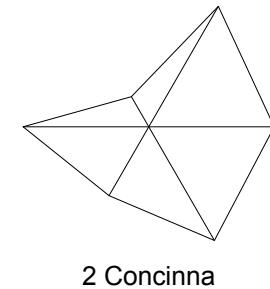
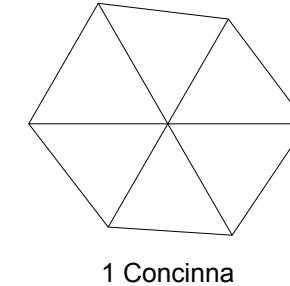
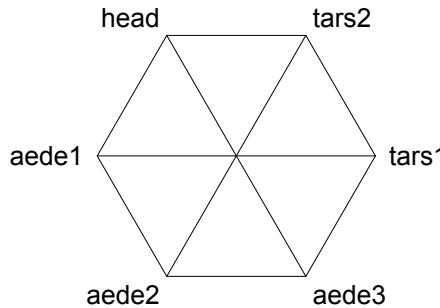
# Tours

Something  
hidden

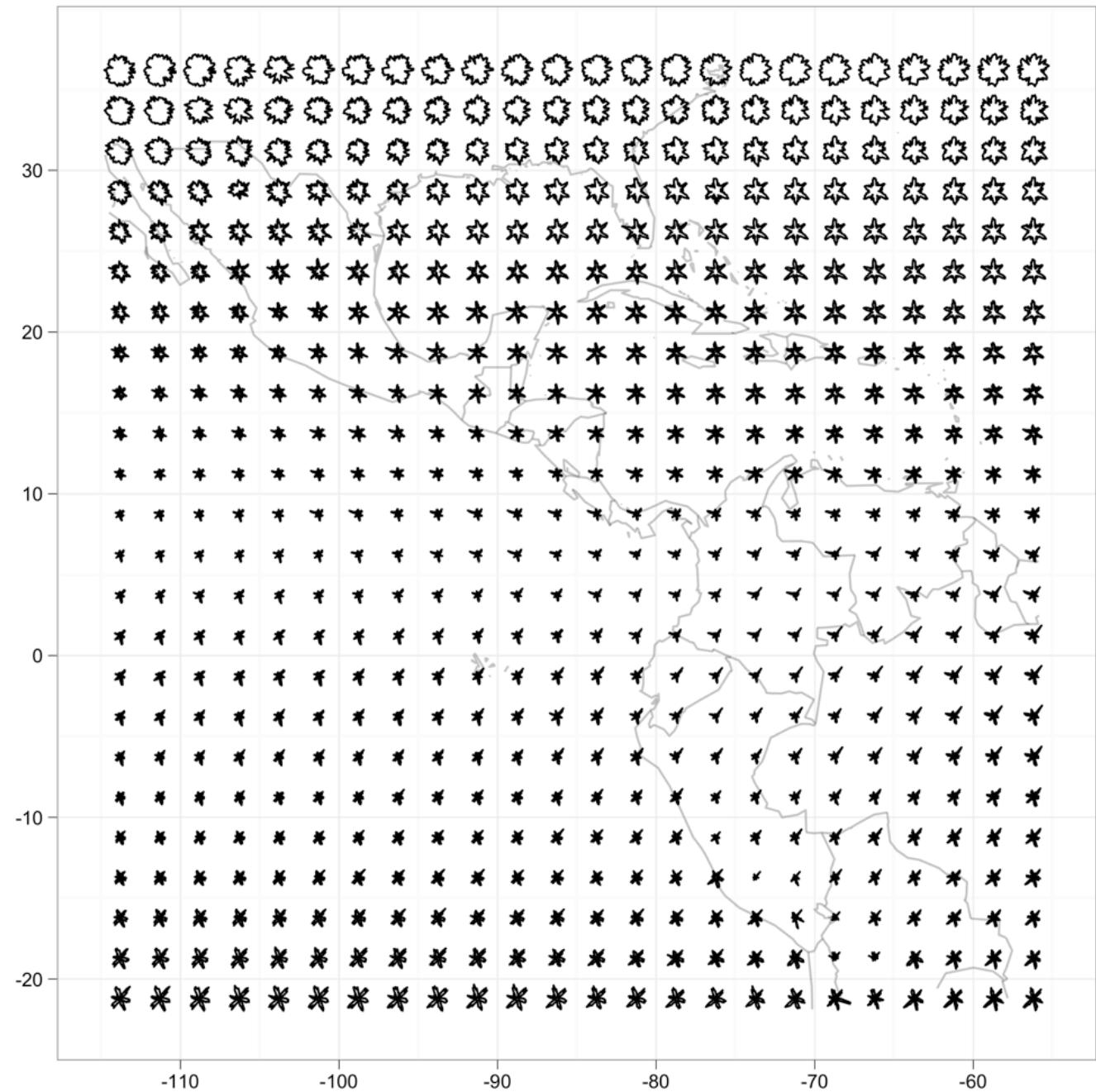


# Icons

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



# Icons



Example:  
Monthly  
ozone for 6  
years over  
central  
America.

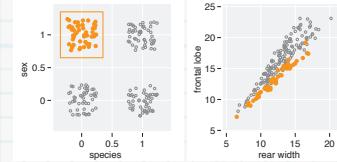
What do you  
see?

# Multiple linked plots

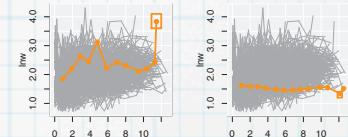
Interactive graphics (direct manipulation graphics) allow the user to brush plot elements and convey this selection to other plots.

# Terminology

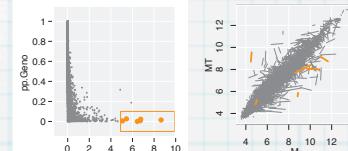
Brushing (one-to-one)



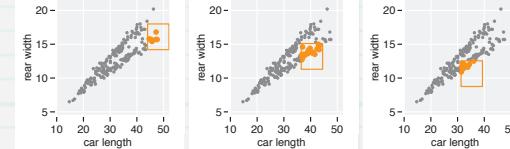
Brushing (categorical variable)



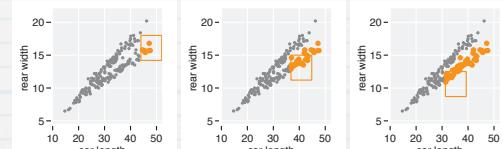
Brushing (one point to one line)



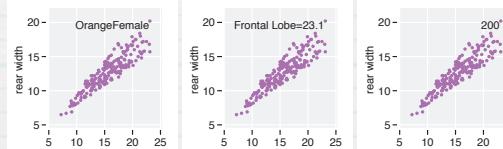
Transient brushing



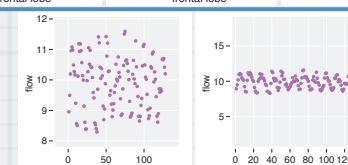
Persistent painting



Identification

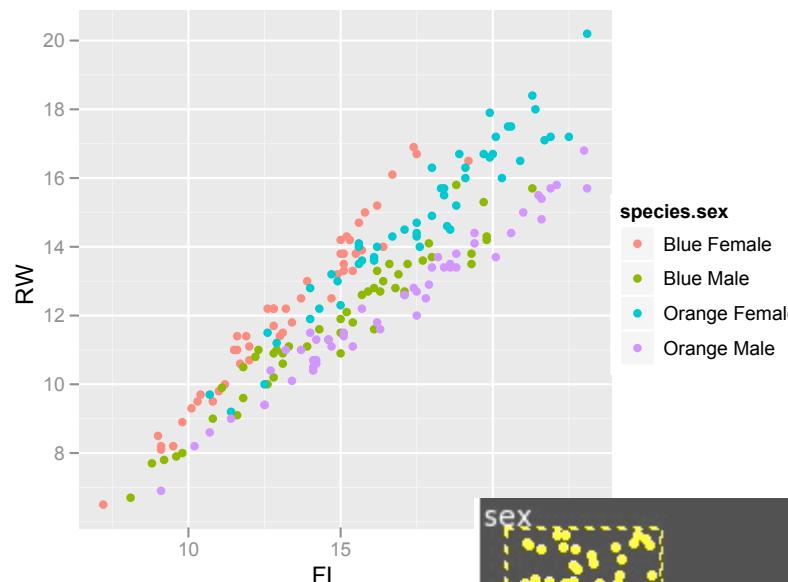


Scaling



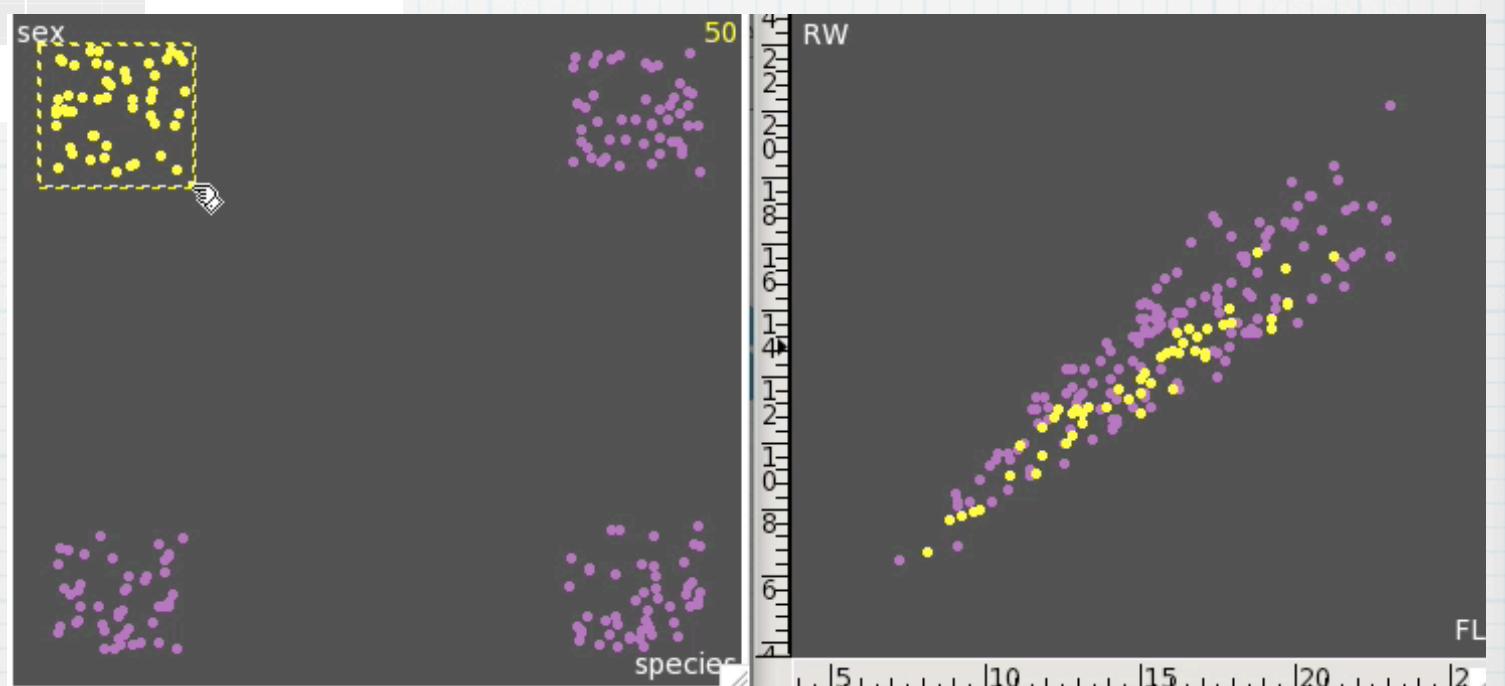
This

# 1-1 Linking



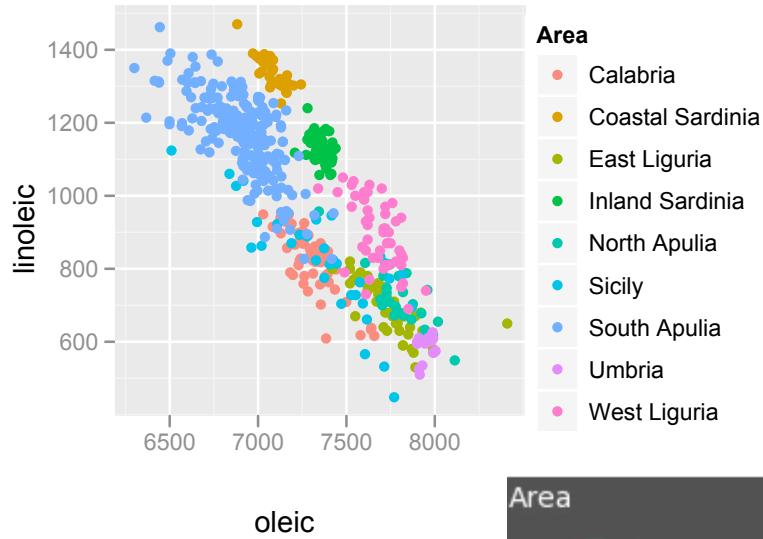
What do you  
learn  
differently?

OR This?



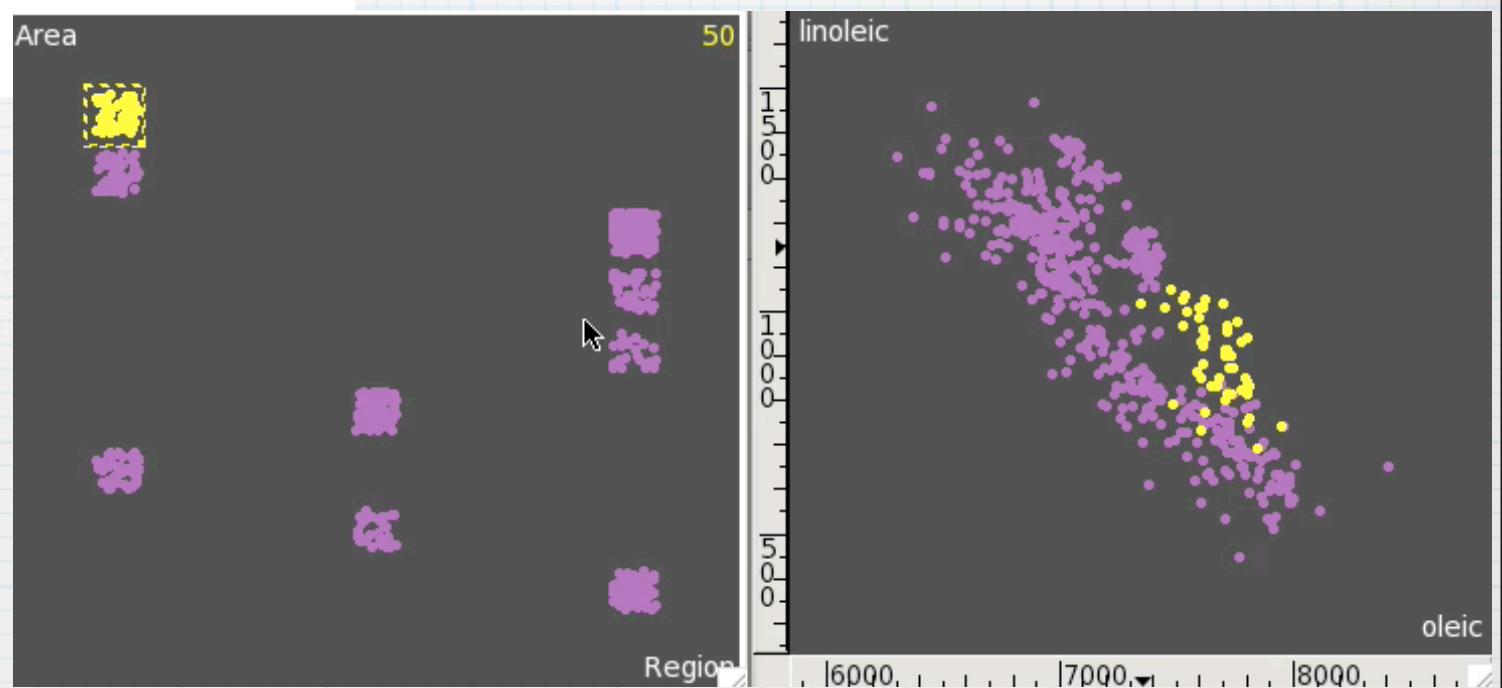
This

# 1-1 Linking



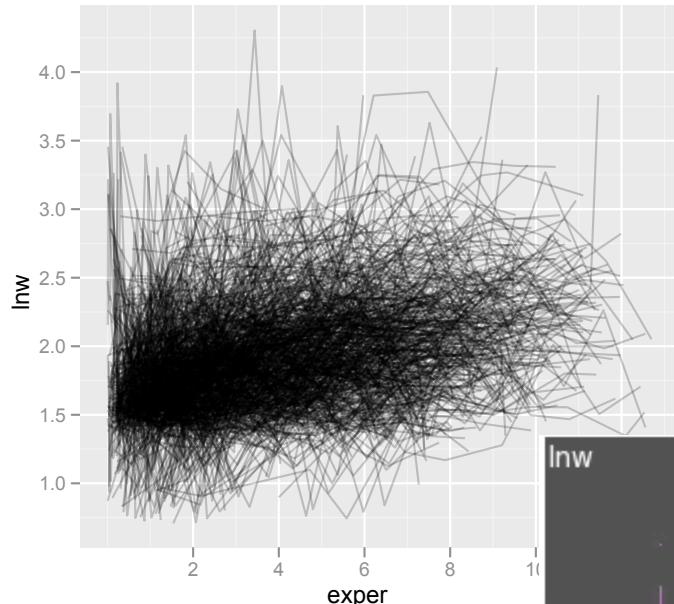
What do you  
learn  
differently?

OR This?



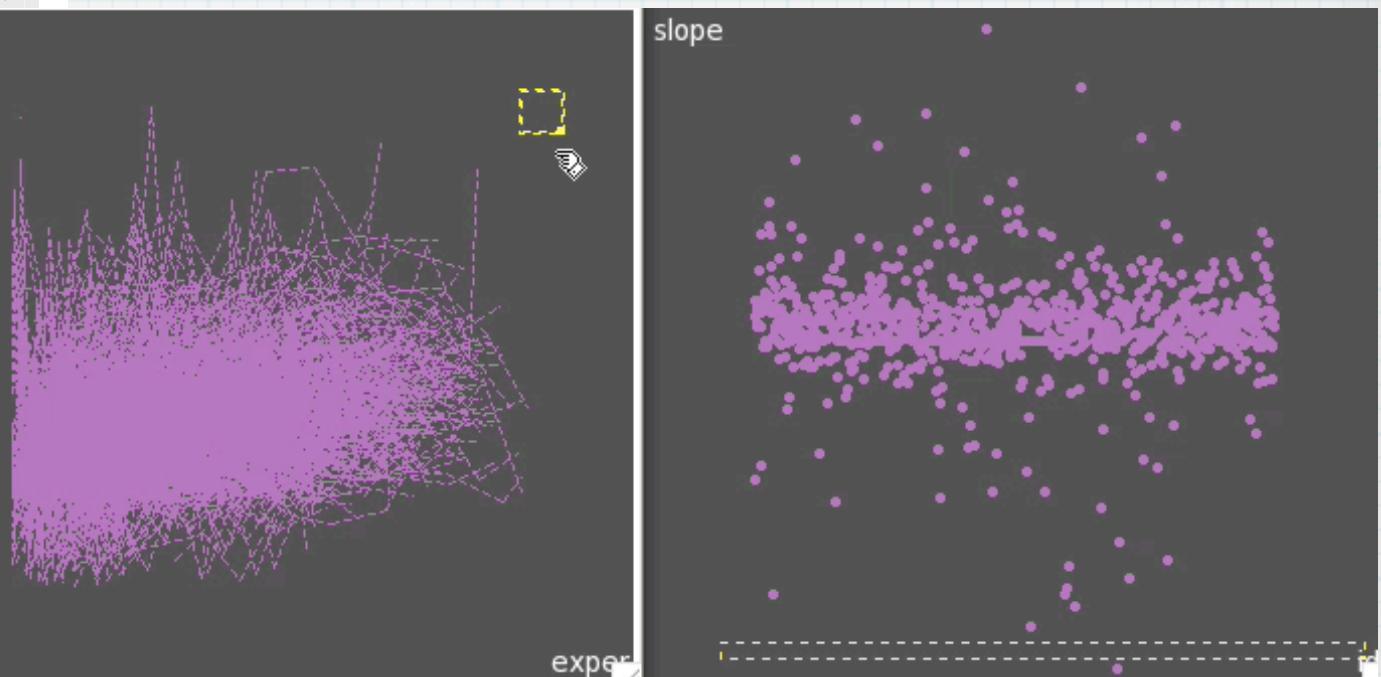
# Categorical Variable Linking

This



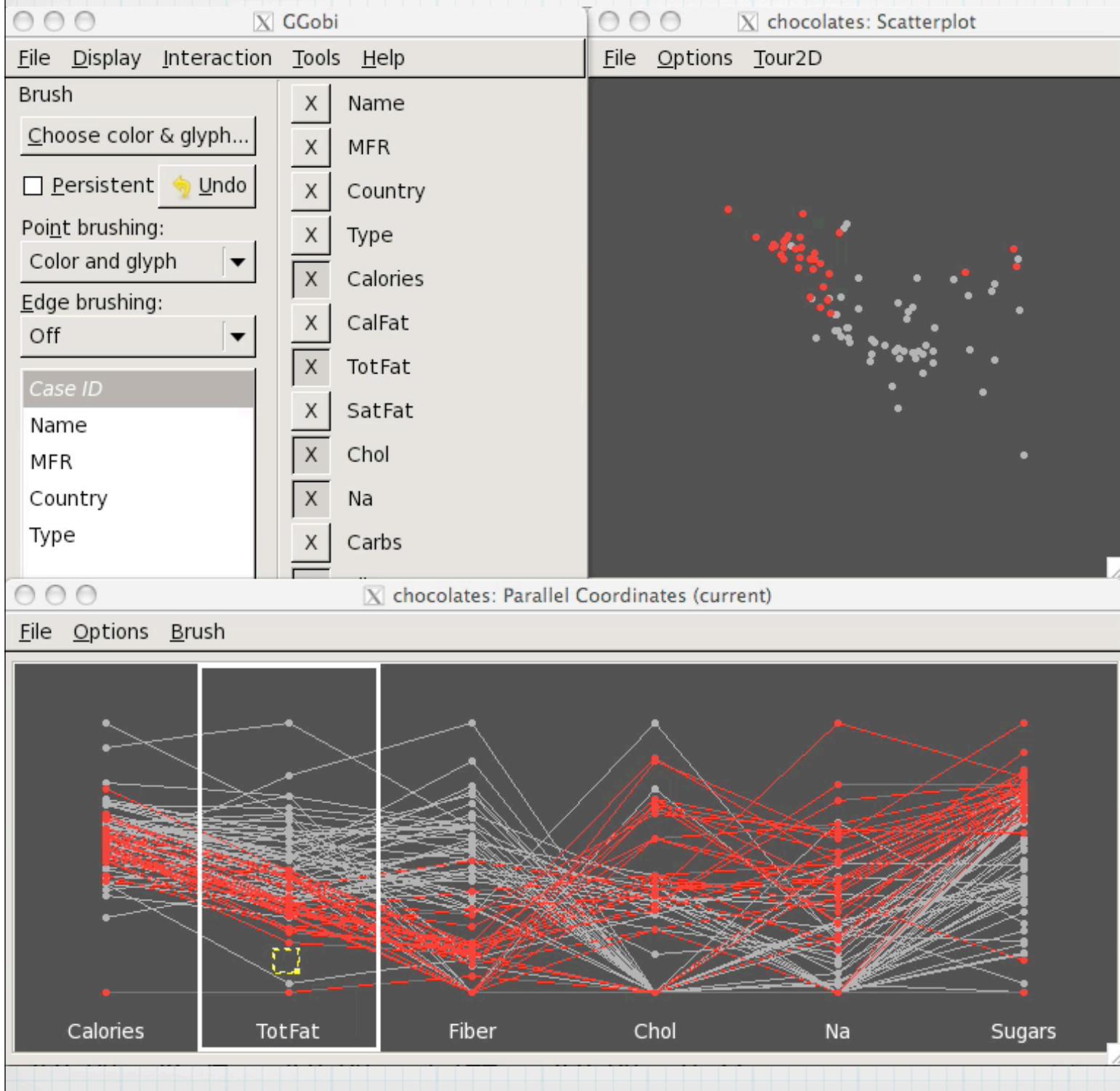
Longitudinal measurements on wages with workforce experience: 888 subjects.

OR This?

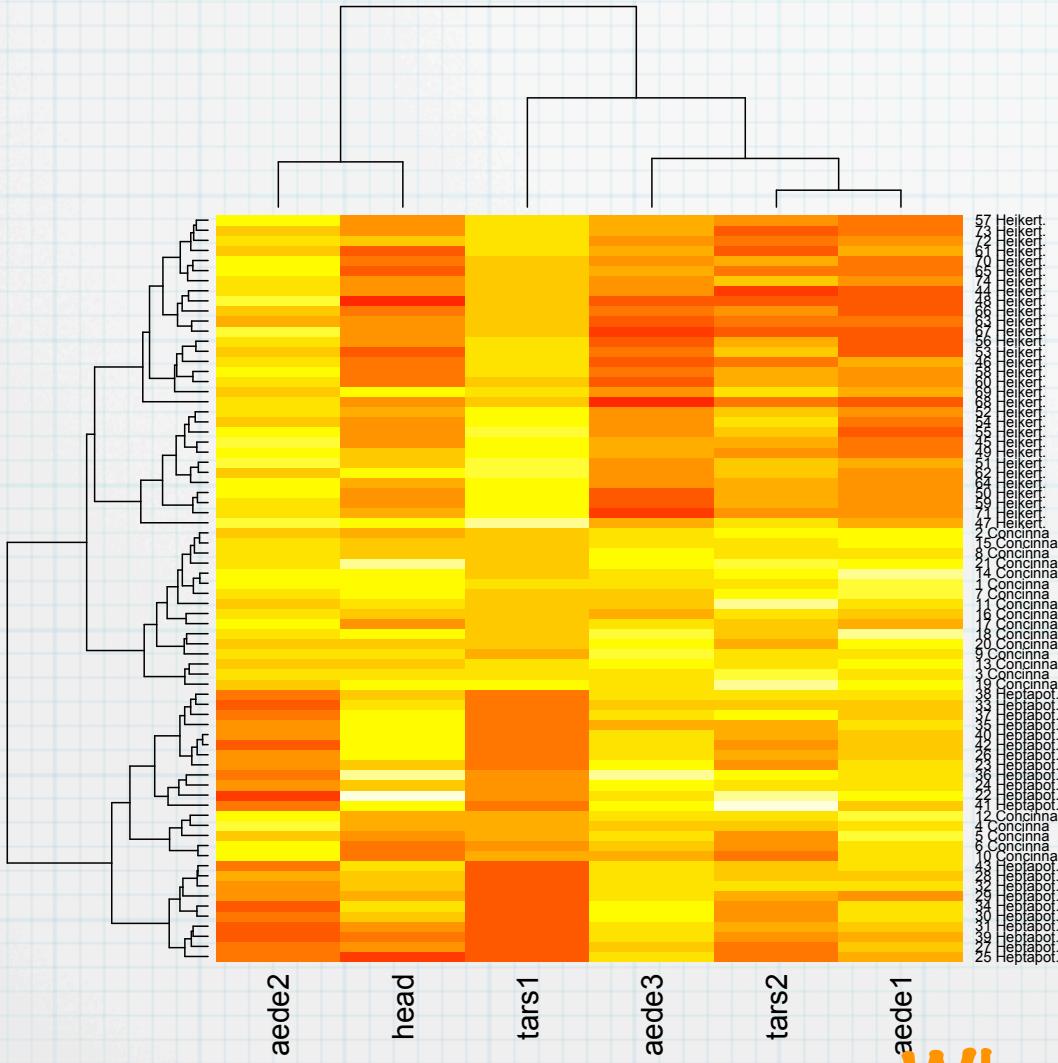


# Example

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?

This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 United States License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/3.0/us/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.