MONASH University

# ETC3250 Business Analytics: Advanced Plotting

**Souhaib Ben Taieb, Di Cook, Rob Hyndman**

September 3, 2015

# Multivariate data?

- 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 (histogram, dotplot, density plot, barchart), then 2 (scatterplot, side-by-side boxplots, . . . ), then more variables.

# Plotting in more than 2-dimensions

- Scatterplot matrix
- Parallel coordinate plot
- Tours

# Scatterplot matrix

- Show all pairs of variables
- If numeric, use a scatterplot - this is the traditional
- If variables types different, use an appropriate plot accordingly

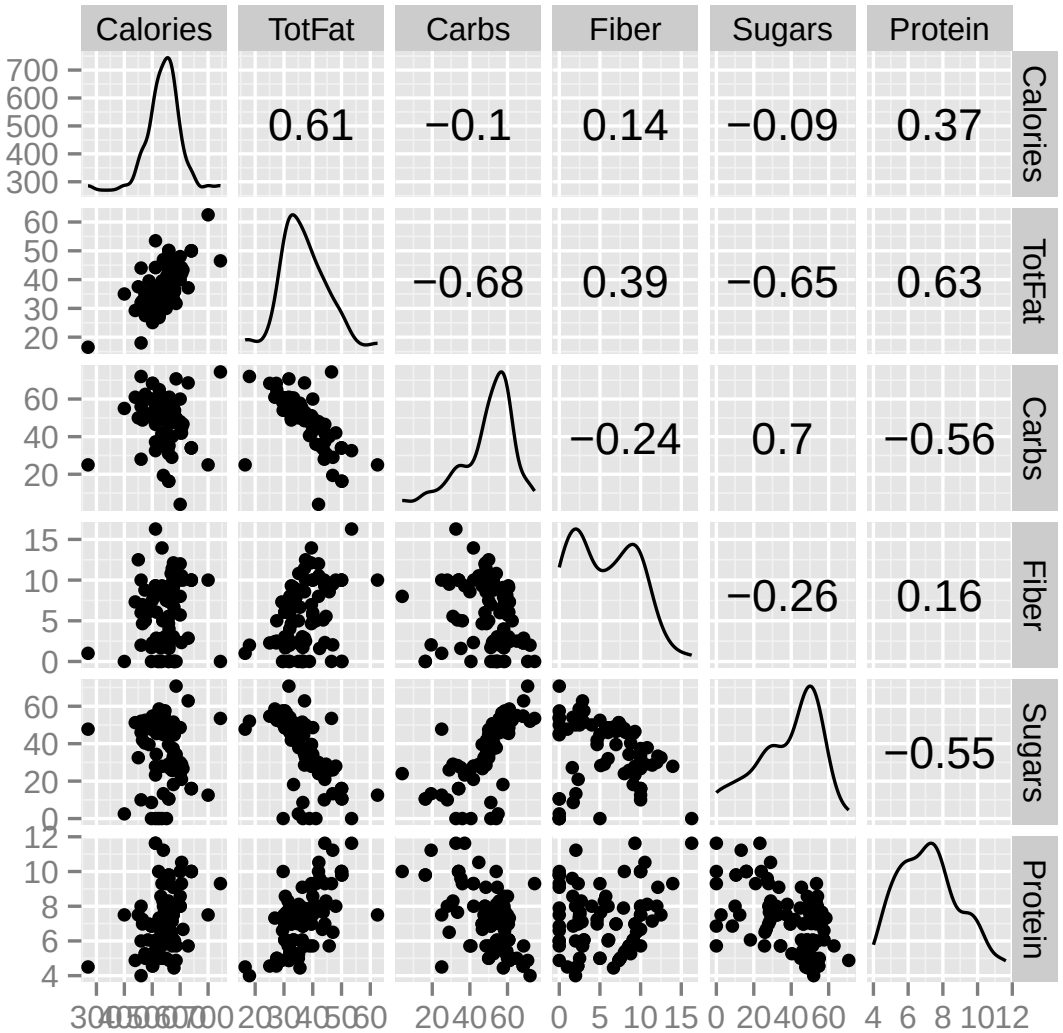
# Example: chocolates nutrition info

- Nutritional information for the equivalent of 100g bars: Calories, CalFat, TotFat, SatFat, Chol, Na, Carbs, Fiber, Sugars, Protein
- 88 chocolates from 30 different manufacturers, 8 different countries
- Both dark and milk chocolates

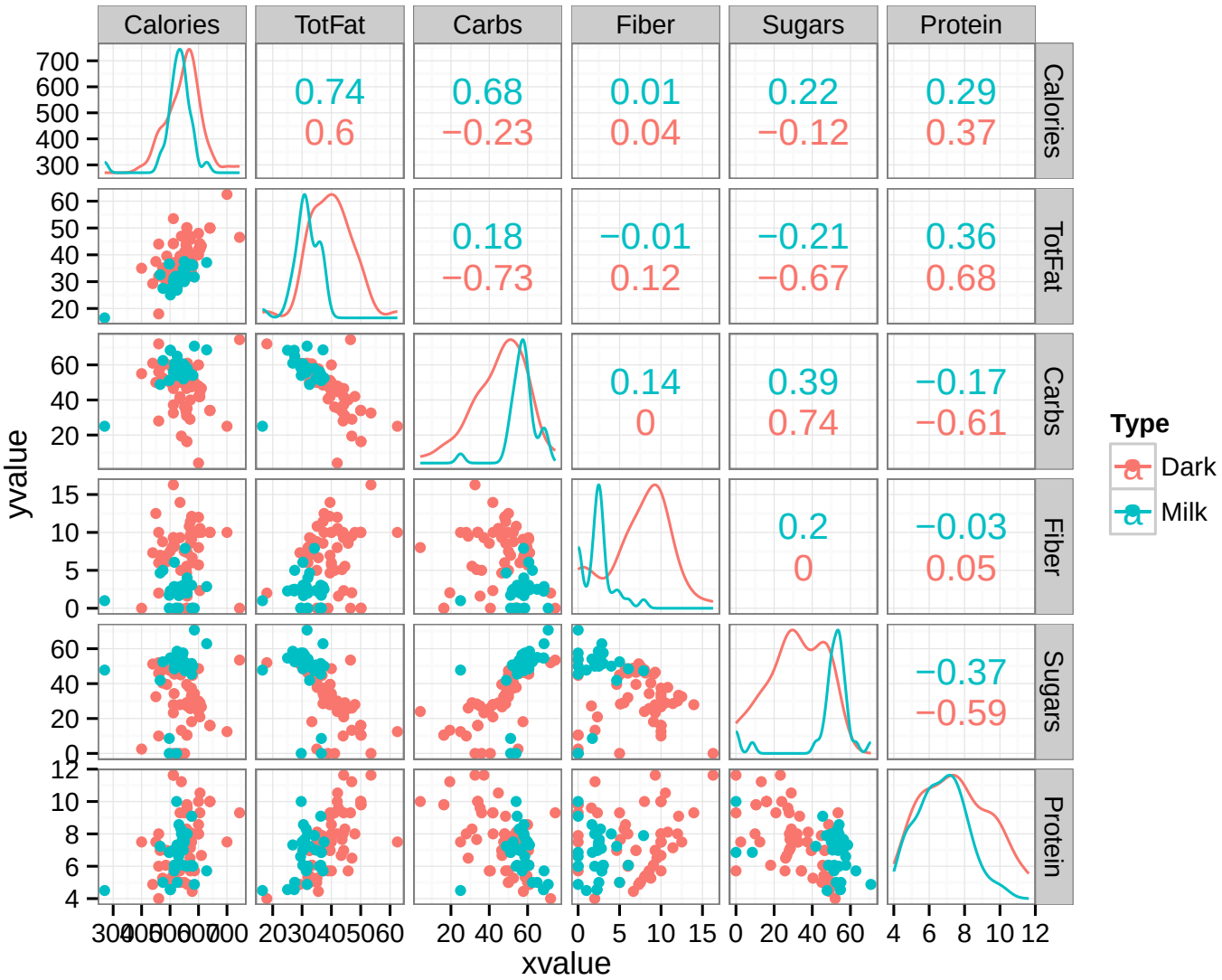
```
## [1] 88 14
```

```
##
## Dark Milk
##   56   32
```

# Scatterplot matrix



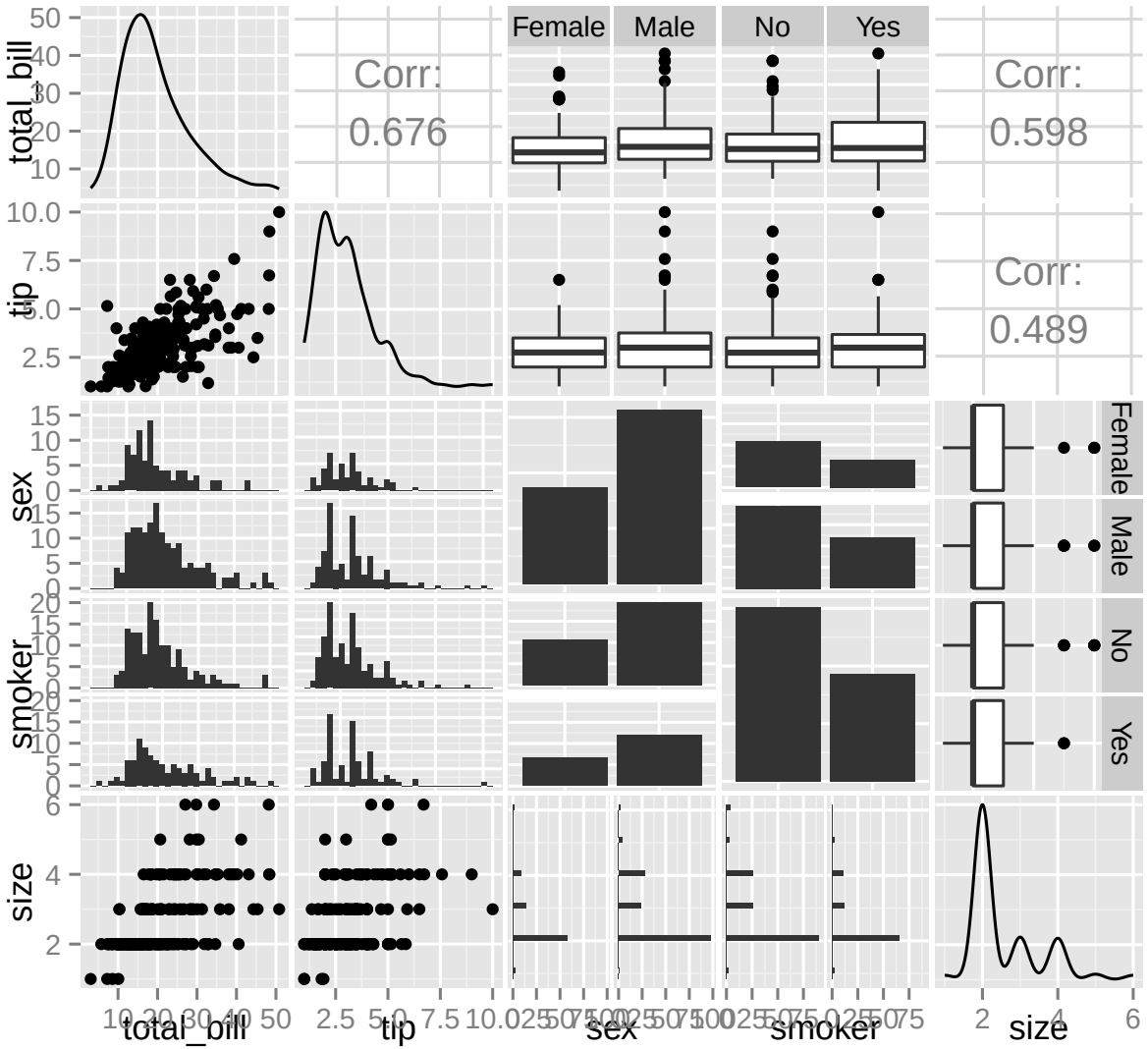
# Colored by type



# Difference

- How do milk and dark chocolates differ?
- What are some other features in the data that are visible?

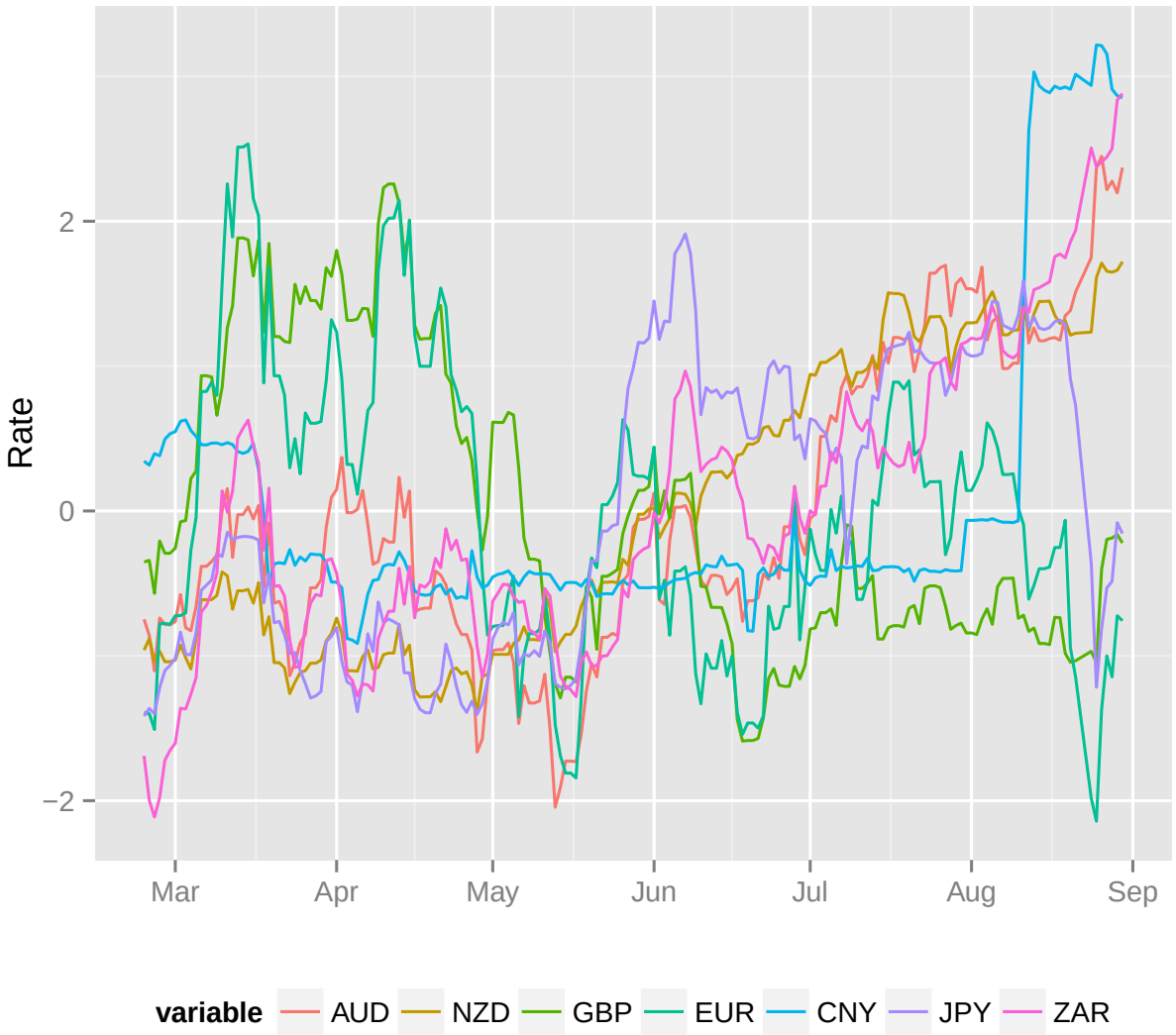
# Catgorical variables



# What do you see?

- postive association between tip and total bill
- bill slightly higher when male is paying, on average
- . . .
- . . .

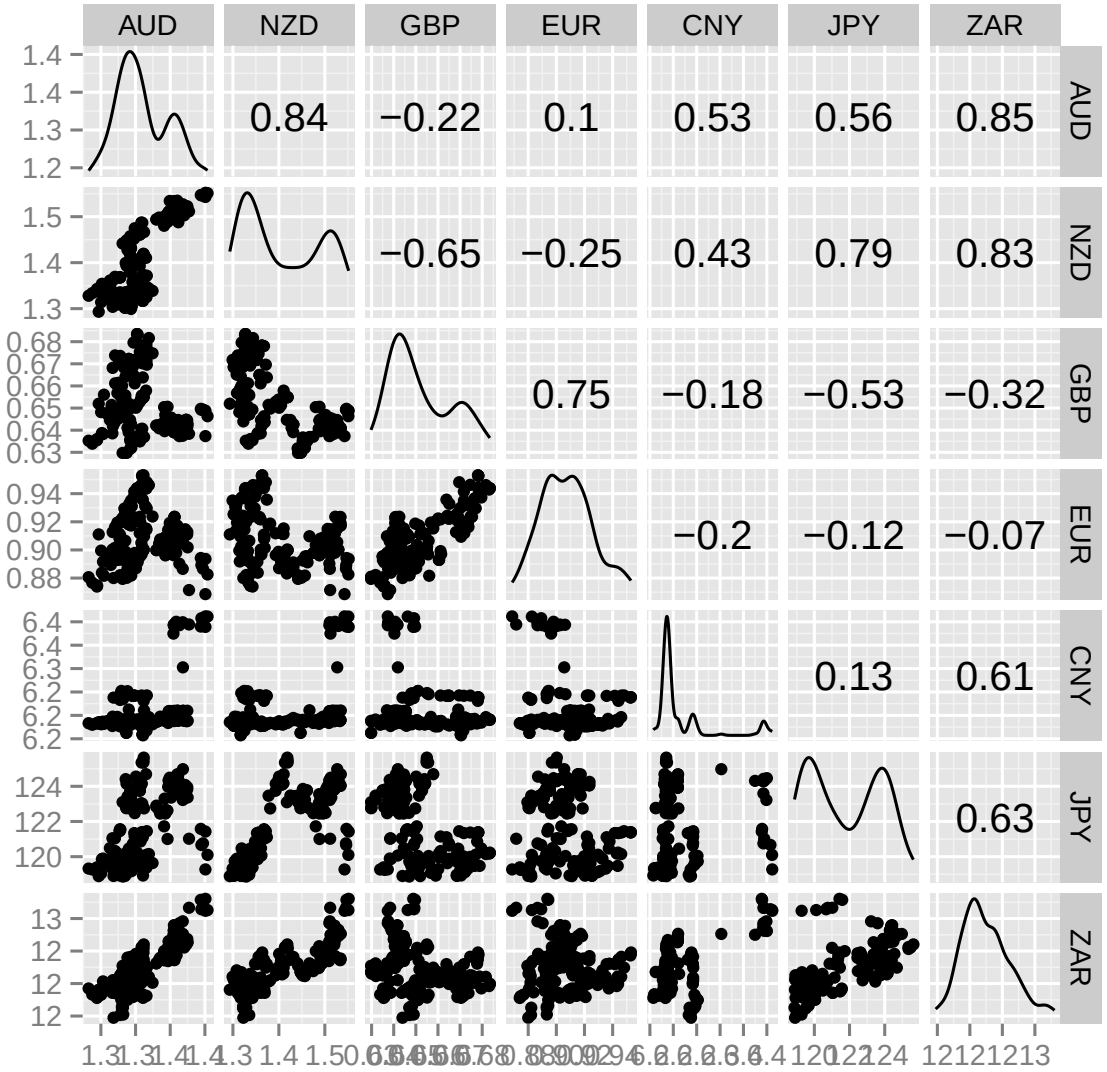
# Example: Exchange rates



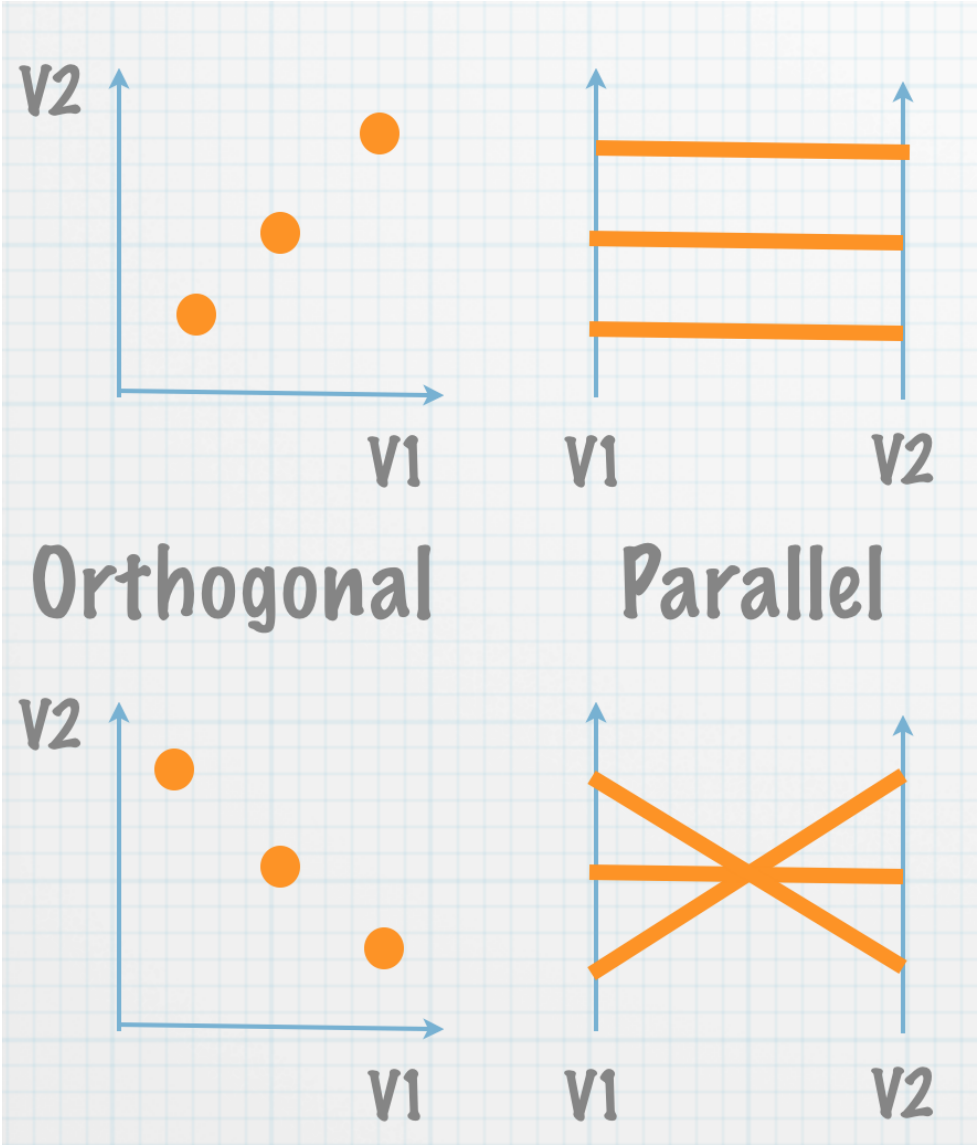
# PCA on exchange rates

```
## Standard deviations:
## [1] 1.97 1.27 0.98 0.50 0.41 0.32 0.14
##
## Rotation:
##        PC1      PC2     PC3     PC4    PC5    PC6    PC7
## AUD   0.43 -0.35049  0.0048  0.5234 -0.100  0.468 -0.439
## NZD   0.49  0.00033 -0.1594  0.2544 -0.191 -0.025  0.794
## GBP  -0.32 -0.56940  0.0850 -0.0908  0.483  0.425  0.376
## EUR  -0.16 -0.68565 -0.2948 -0.1158 -0.496 -0.393 -0.055
## CNY   0.29 -0.15128  0.7657 -0.4469 -0.309  0.085  0.043
## JPY   0.39  0.02537 -0.5343 -0.6633  0.071  0.318 -0.111
## ZAR   0.45 -0.24349  0.0928 -0.0059  0.611 -0.580 -0.131
```

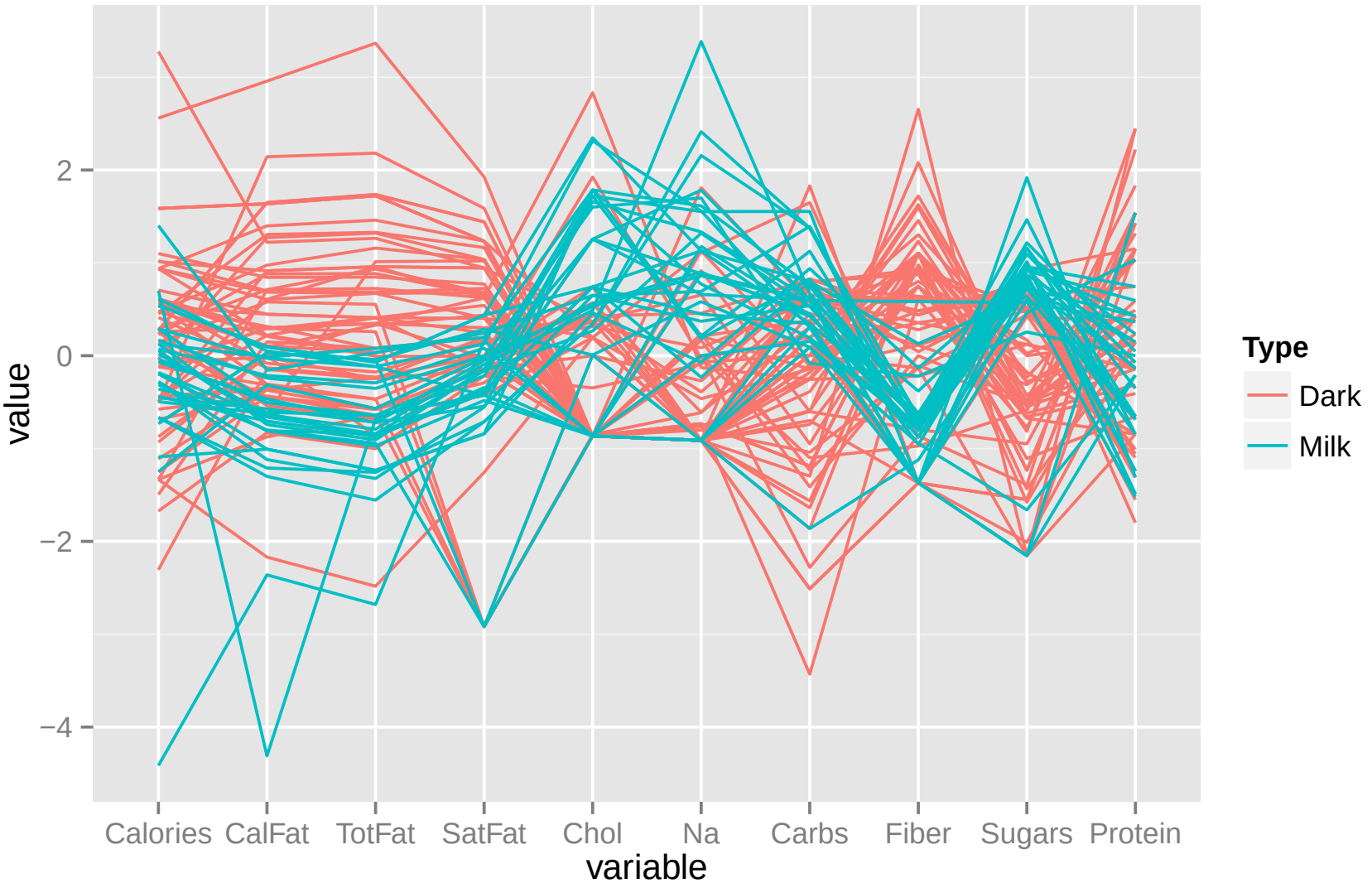
# Example: Exchange rates



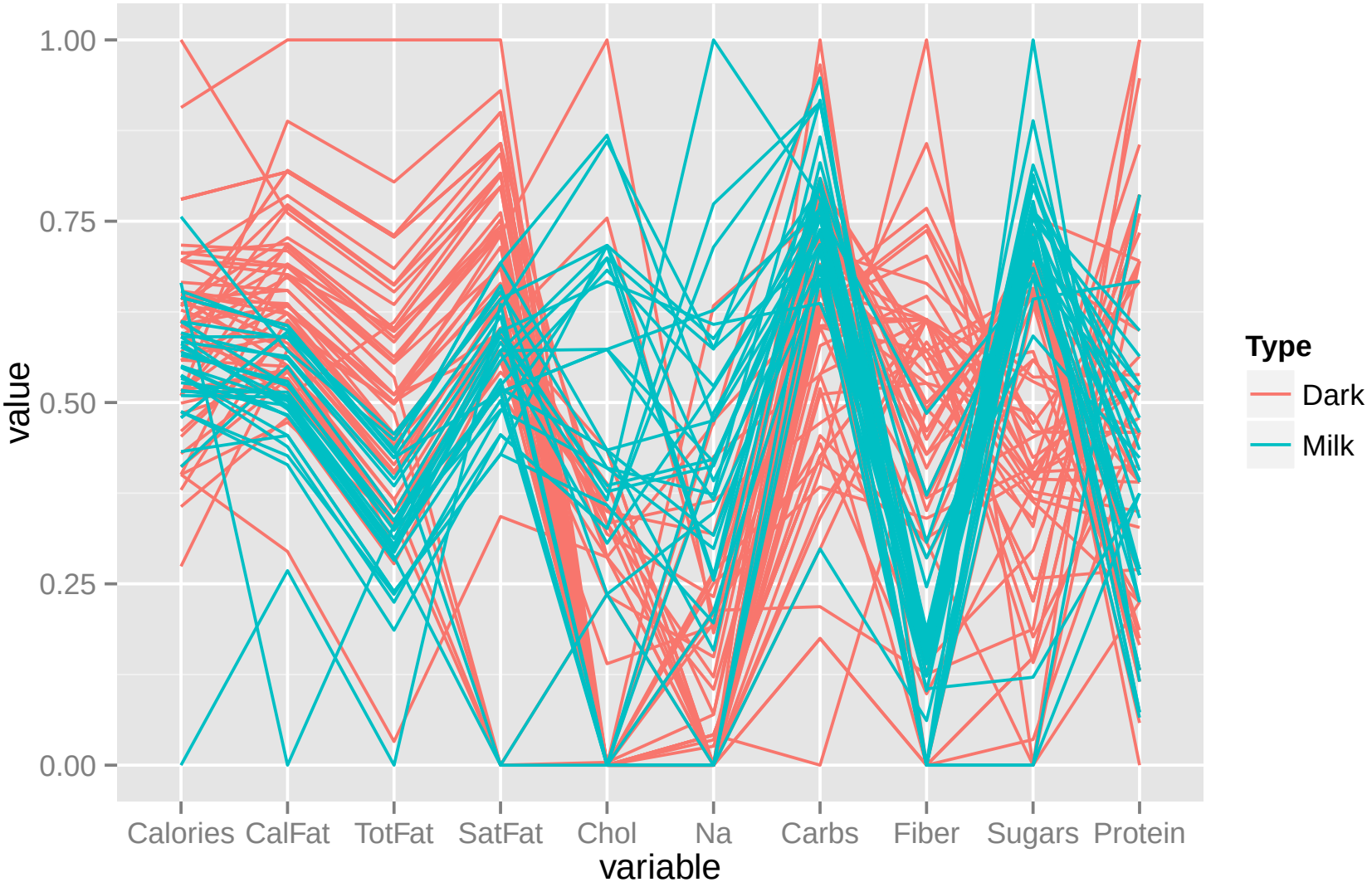
# Parallel coordinate plot



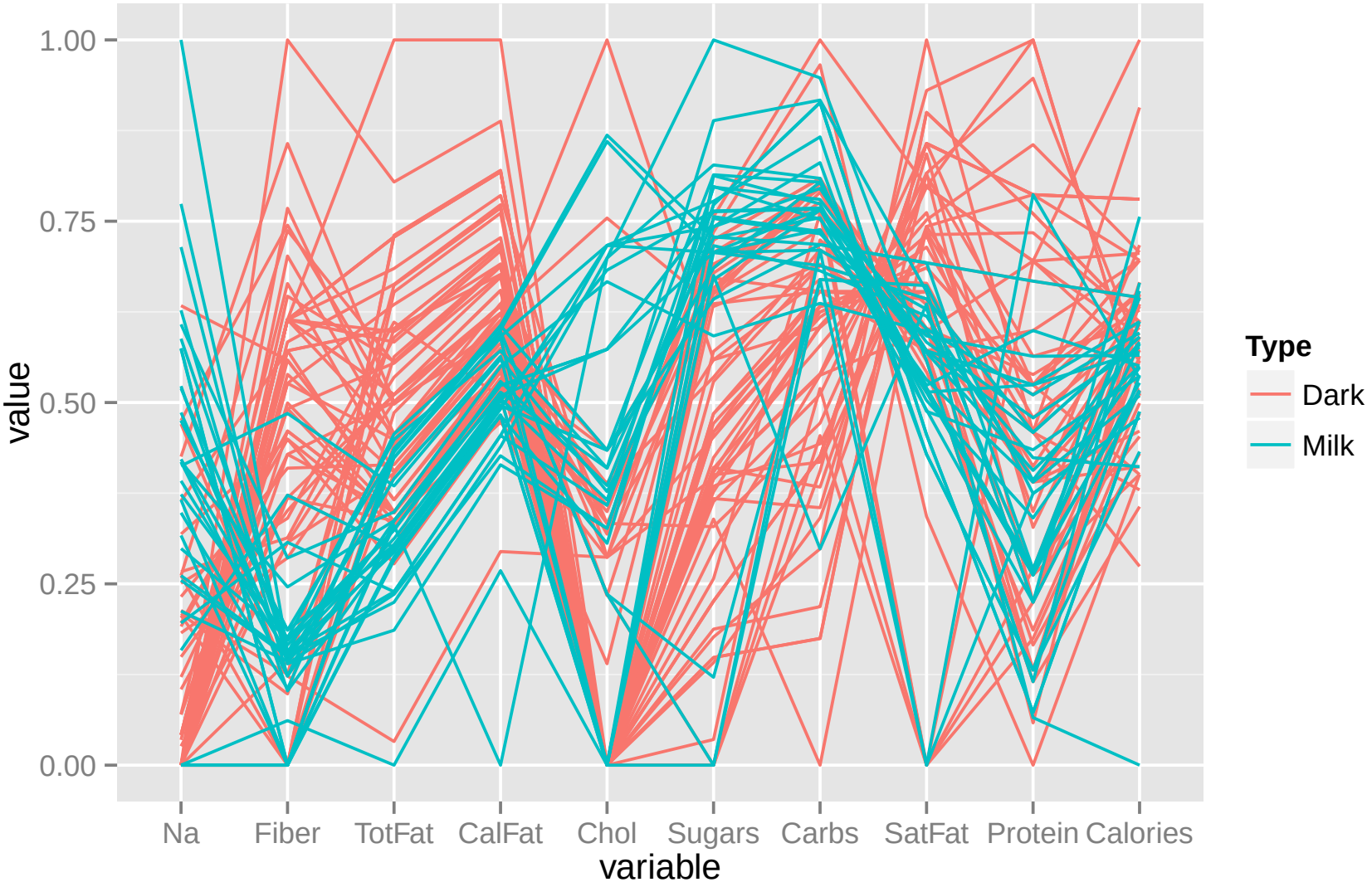
# Example: chocolates



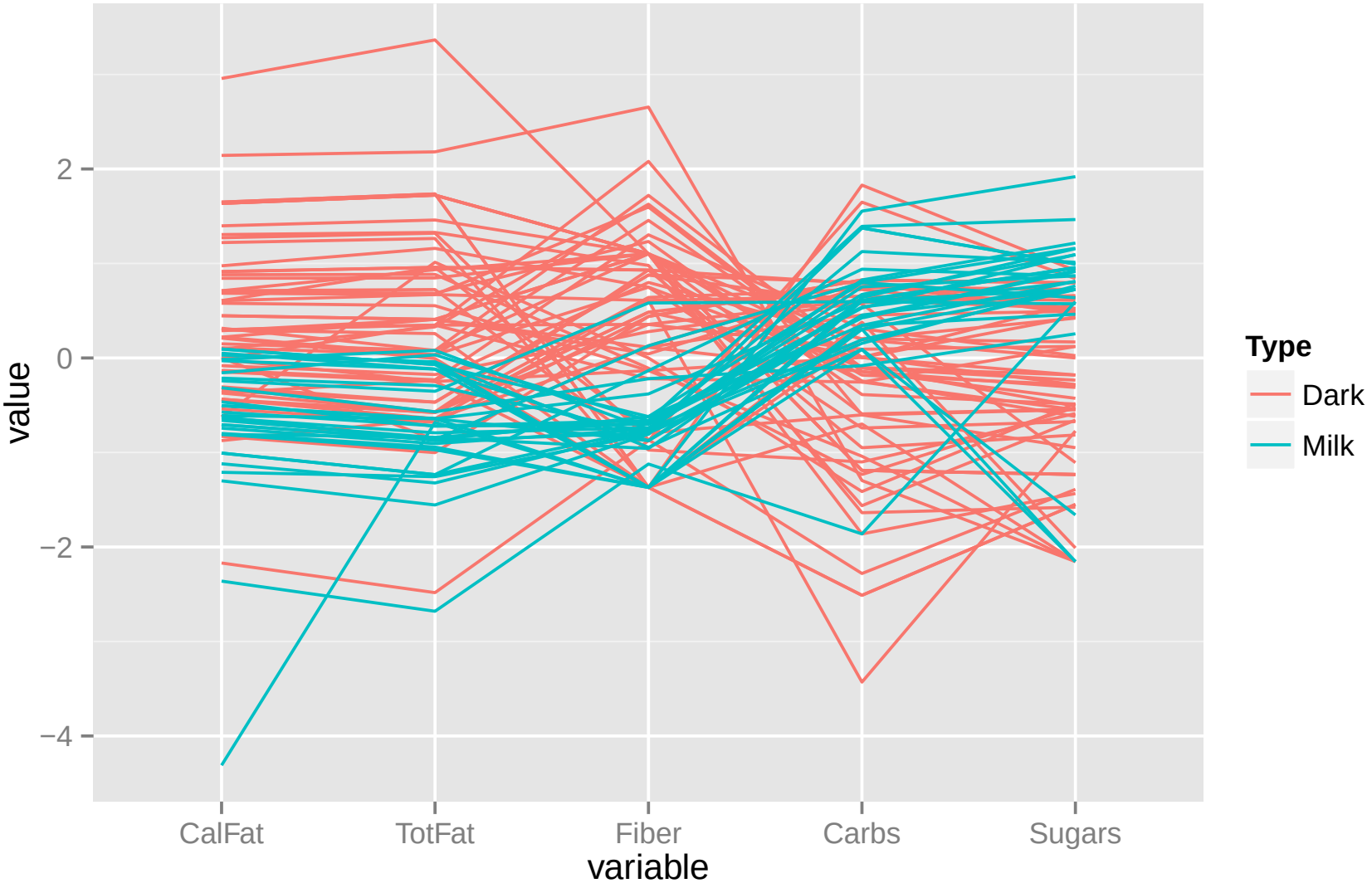
# Scaling changes appearance



# Re-ordering variables can help more



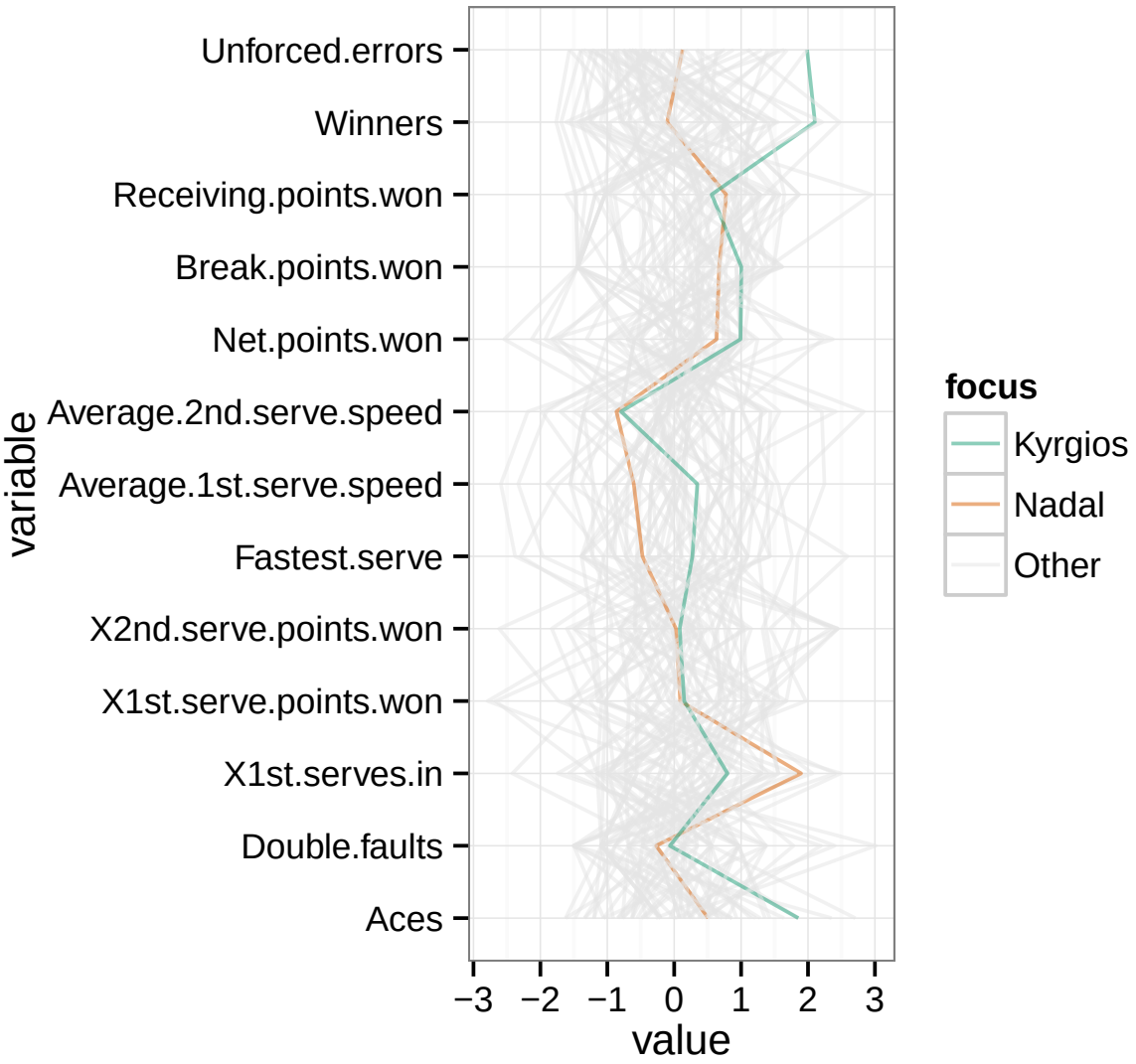
# Re-ordering variables can help more



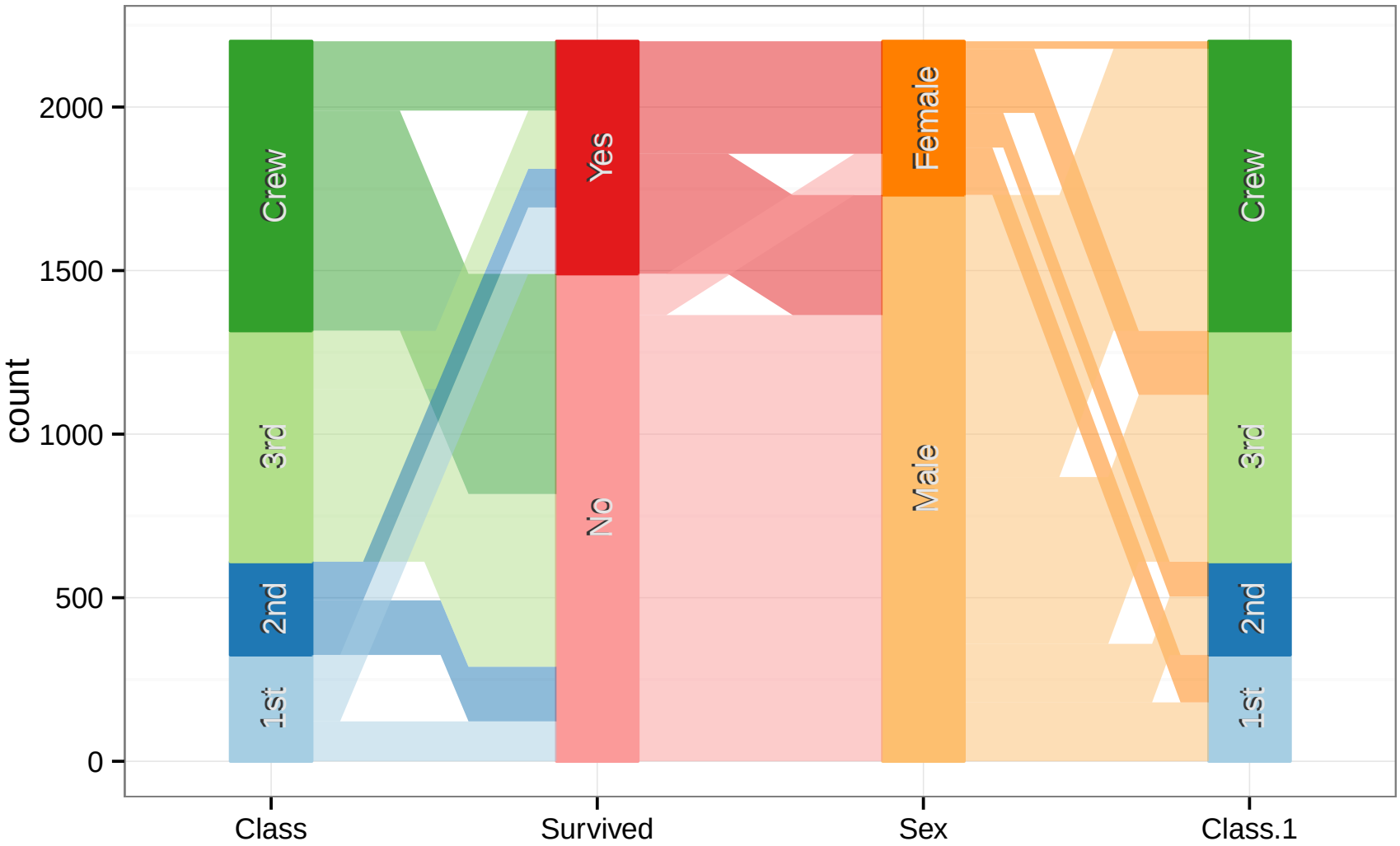
# What do we learn?

- difference between milk and dark
- unusual chocolates
-

# Example: Men's Tennis, Wimbledon



# Categorical variables



# Tours

- Let's take a look at the chocolates with a tour, before I explain what the method does.

```
library(rggobi)
ggobi(choc)
# Run outside of R
ggobi chocolates.csv
```

# Tour explanation

- 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.

- Changing projection matrix values produces the movie.

# Tour

- Explanation of the tour
- Note how similar this is to biplot

# Tours showing simple data structure

Click on the links to see the videos on vimeo

- Clusters: clusters
- Nonlinear dependence: nonlinear
- Nonlinear+noise: nonlinear w/noise
- Labelled classes, outliers: supervised classification
- Multivariate normal: Pollen data

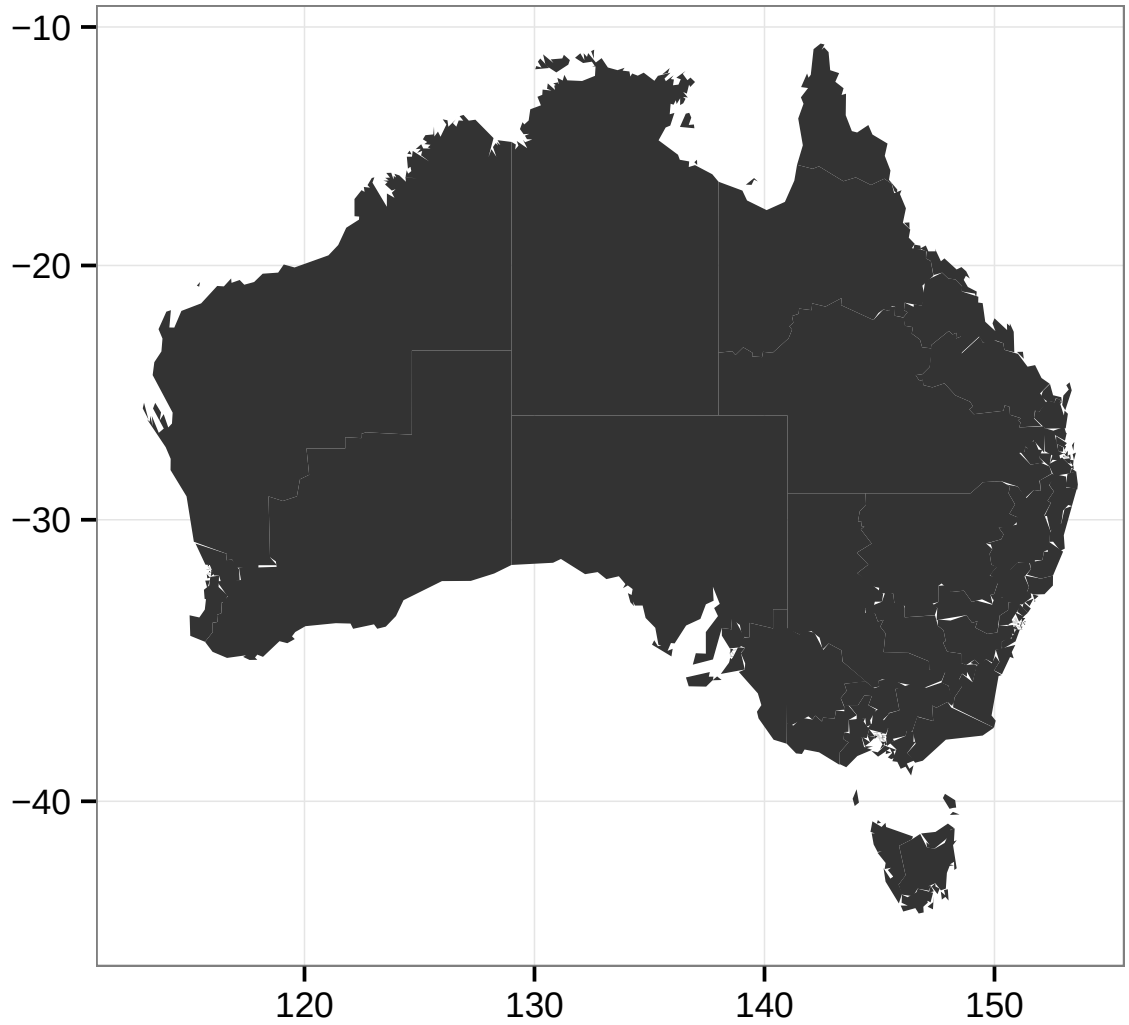
# Linking on NRC data

- Putting all of these plots together, and linking them enables exploring for all sorts of structure in the data
- The video here illustrates this approach for exploring the National Research Council rankings of Statistics graduate programs in the USA.
- The data contains two different ranking schemes (R, S) both of which rely on surveys of faculty on how other departments are viewed. (Rough explanation) One creates the ranks using regression of the faculty ranking against information collected about each department, and then predicts the ranks. The second is based on a factor analysis of the department information.
- Statistics collected include number of students, faculty, female faculty and students, number of professors in each rank, time to degree, average GREs of incoming students, average number of publications/citations per faculty, student activities, groups, ...

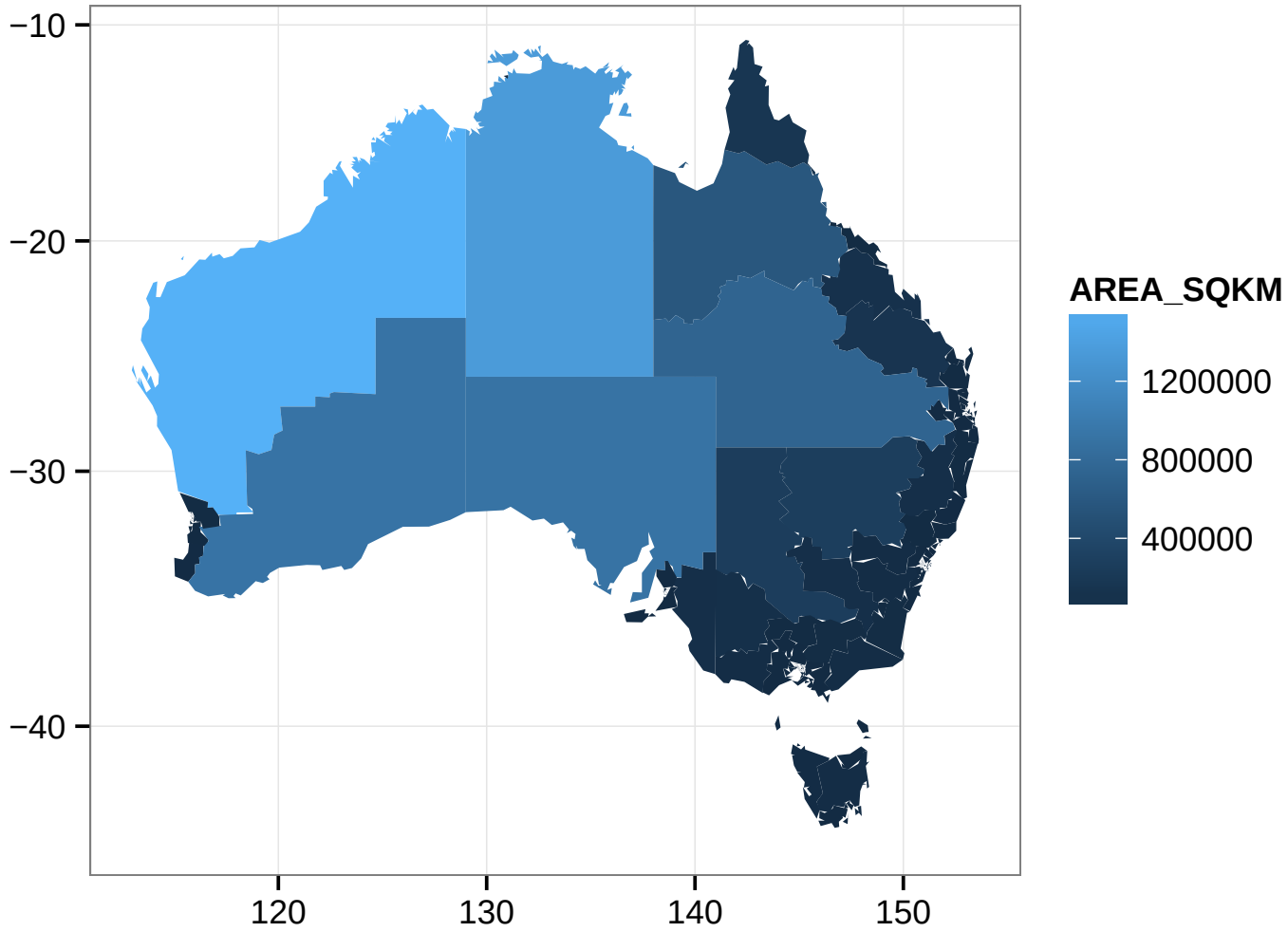
# Maps: Australia electoral boundaries

Boundaries come from

`http://www.aec.gov.au/electorates/maps.htm`

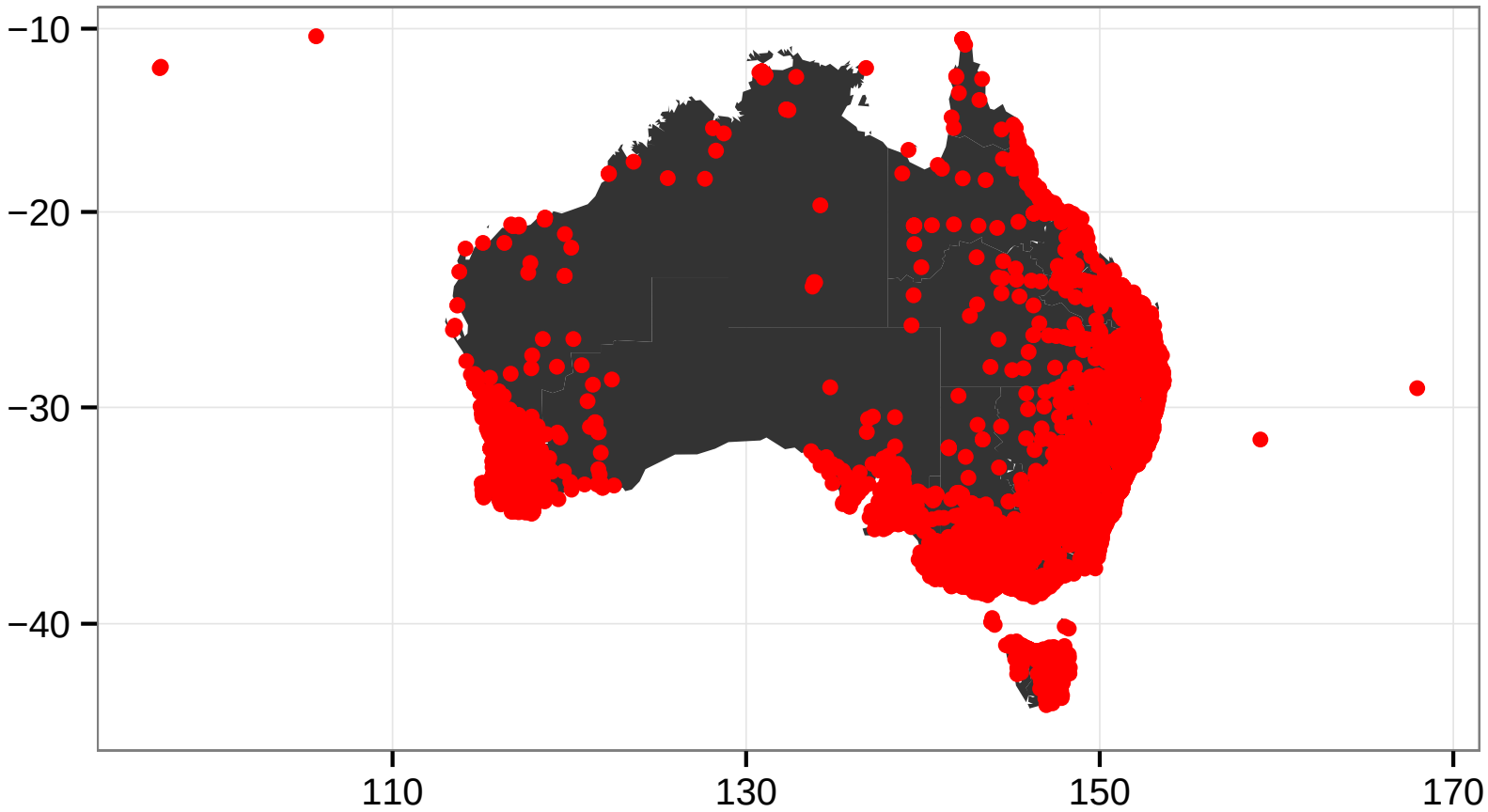


# Make it a chloropleth map



# Add locations of polling stations

```
##
##   ACT   NSW    NT   QLD    SA   TAS   VIC    WA
##    85  2740    50  1482   703   333  2062   811
```



# Last question

Is a pie chart simply a bar chart in polar coordinates?

Can you use the grammar of graphics of ggplot2 to answer this question?