Graphics in R: No matter how big, we can make it small

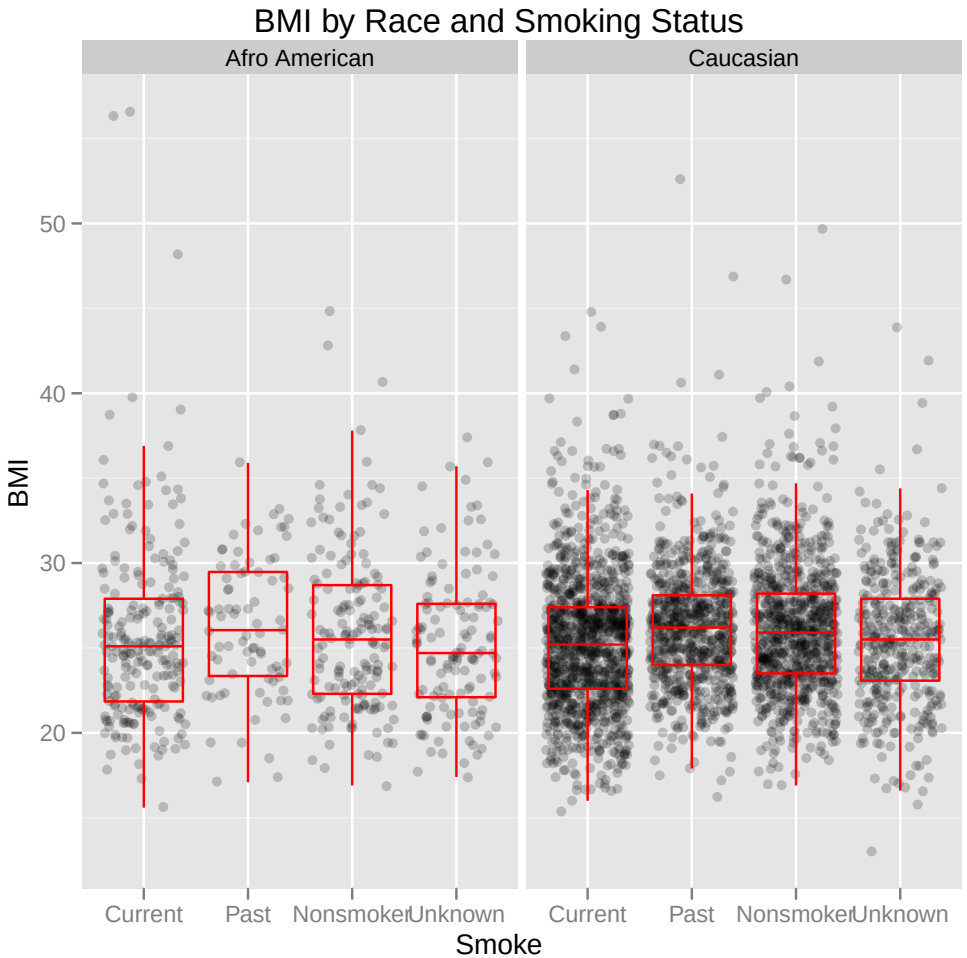
Nicholas Lewin-Koh

Genentech

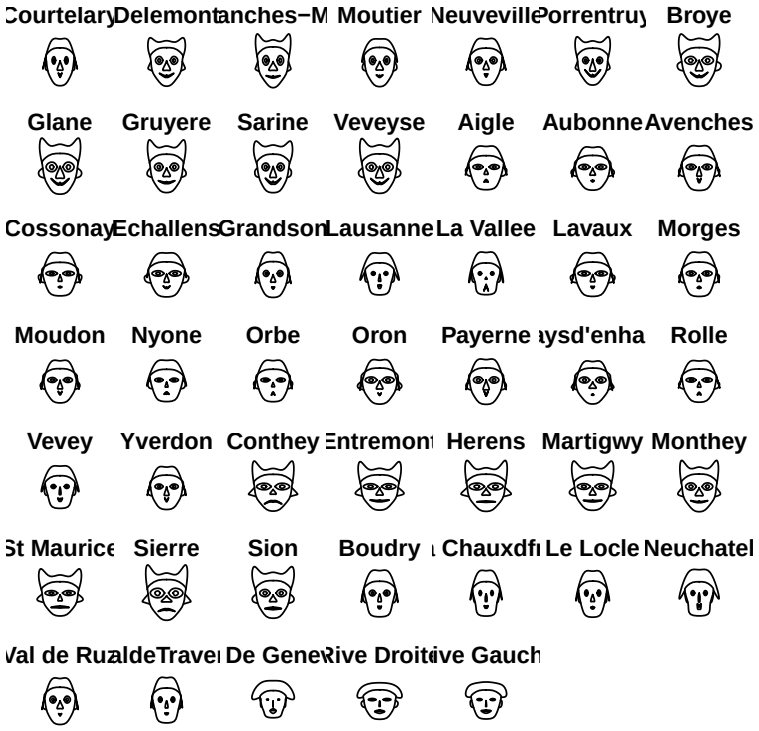
June 11, 2009

Outline

Kaleidoscope

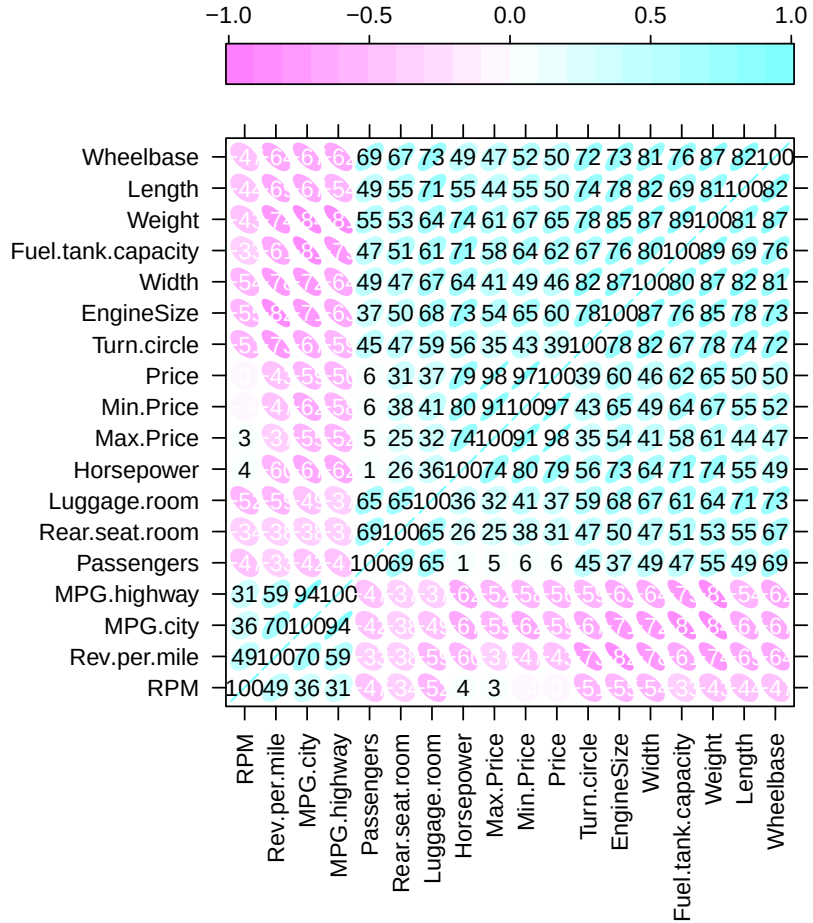


Kaleidoscope



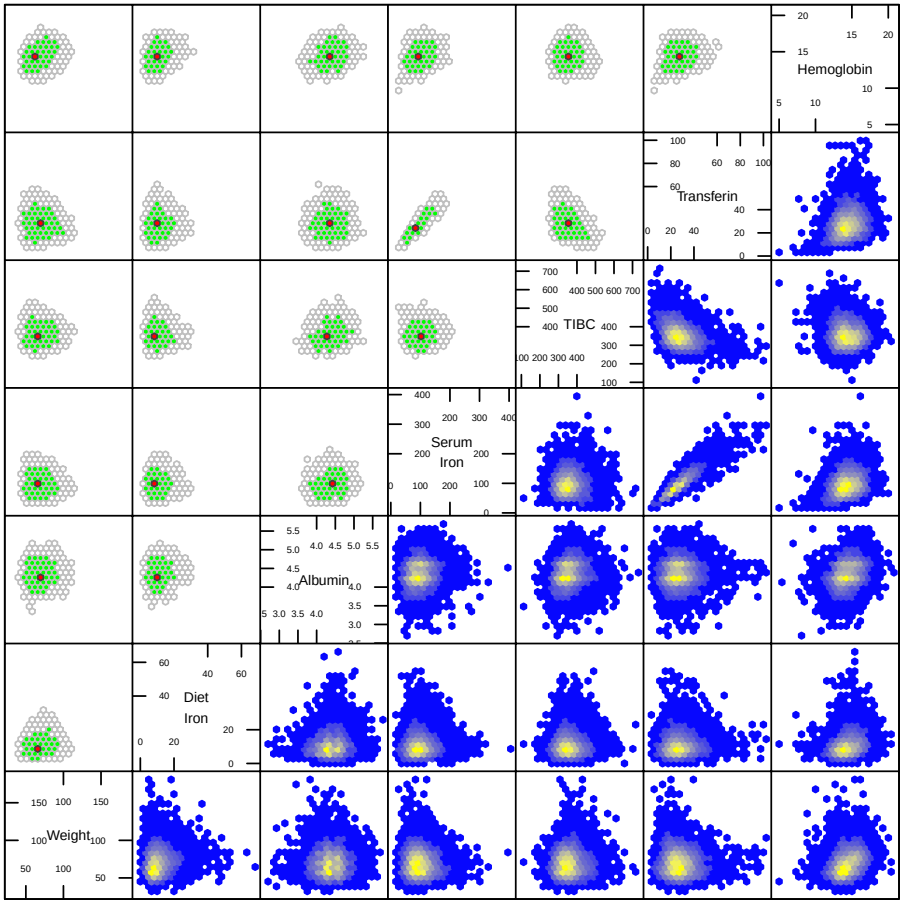
Chernoff faces, 1888 Swiss, French speaking provinces. function from aplpack, by H. P. Wolf

Kaleidoscope



Correlation plot of 1993 Cars data, plot code by Deepayan Sarkar

Kaleidoscope



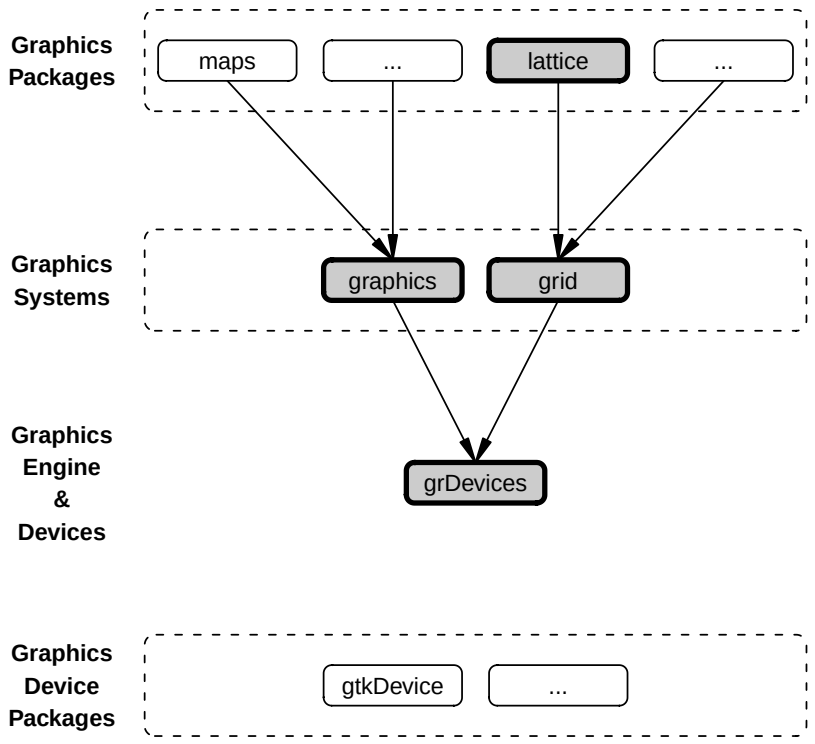
Scatter Plot Matrix

The R graphics system

R's graphic system is composed of:

- High level plotting functions, often in packages such as lattice and ggplot2
- Low level graphics functions, to draw lines, shapes, glyphs, and fills, two systems:
  - Base graphics
  - Grid
- The underlying graphics engine and or device drivers

The R graphics system: Under the hood



What does a graphics system do?

The goal of the graphics system is to create coordinates for each graphical object and render them to a device or canvas. In addition the system might

- Maintain a stack of graphics objects
- Maintain local information about state
- Use information from the stack to manage redrawing and resizing
- ...

Both grid and base graphics do this to varying degrees.

Trade off between grid and base graphics

Grid maintains a much more detailed stack of graphical objects, called grobs. As a result in grid,

- Resizing maintains relative positions and aspect ratios in a more consistent way
- More conrol of redrawing semantics and partioning the device
- Rendering is slower (Maybe, but things have changed since I last looked).

Timings for rendering a simple plot

```
times<-list()
j<-1
for(i in c(100,1000,10000,100000)){
    x<-rnorm(i)
    y<-rnorm(i)
    times[[j]]<-system.time(plot(x,y))
    j<-j+1
    times[[j]]<-system.time(print(xyplot(x~y, alpha=.1)))
    j<-j+1
}
```

```
> times<-list()
> j<-1
```

Timings for rendering a simple plot

No. of Points	Base Graphics	grid (Lattice)
100	0.016	0.112
1,000	0.108	0.208
10,000	1.012	1.108
100,000	10.313	9.989

Table: Comparing a simple xyplot and plot for different size point sets

Surprisingly, at very large numbers of points, grid is outperforming base

Base graphics

As alluded to before the base graphics system is the original system inherited from the S language.

- It is possible to do very nice plots using base graphics
- Many of the default plots on objects in R are done in base graphics, so plot(object) can give a quick view.
- For very fine tuned graphics, where the display needs to be finely partitioned, grid is more appropriate.

gclus

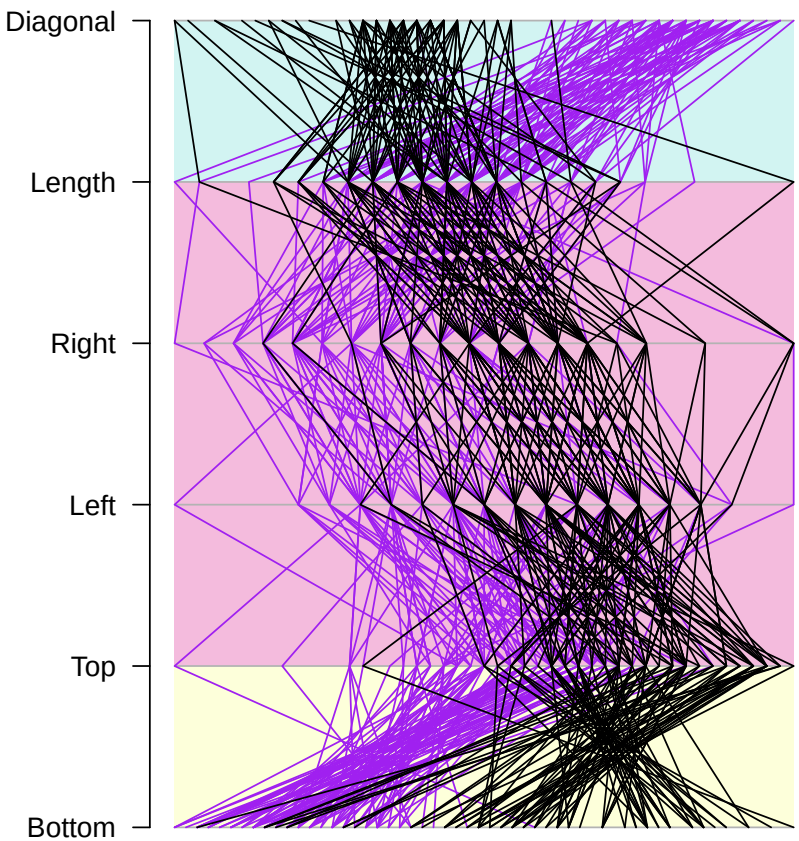
The gclus package, created by Catherine Hurley presents a nice framework for displaying multivariate data.

- Focuses on ordering of plots based on merit indicies.
- **The ordering of the variables (or facets) inmultivariate or conditioned data is more important than any eye candy that goes on the plot**
- **Package presents alternte versions of pairs and paarcoord to emphasize the ordering of the data.**

Swiss Banknotes

Data from:
"Multivariate Statistics A practical approach", by Bernhard Flury and Hans Riedwyl, Chapman and Hall, 1988, Tables 1.1 and 1.2 pp. 5-8.
Six measurements made on 100 genuine Swiss banknotes and 100 counterfeit ones.

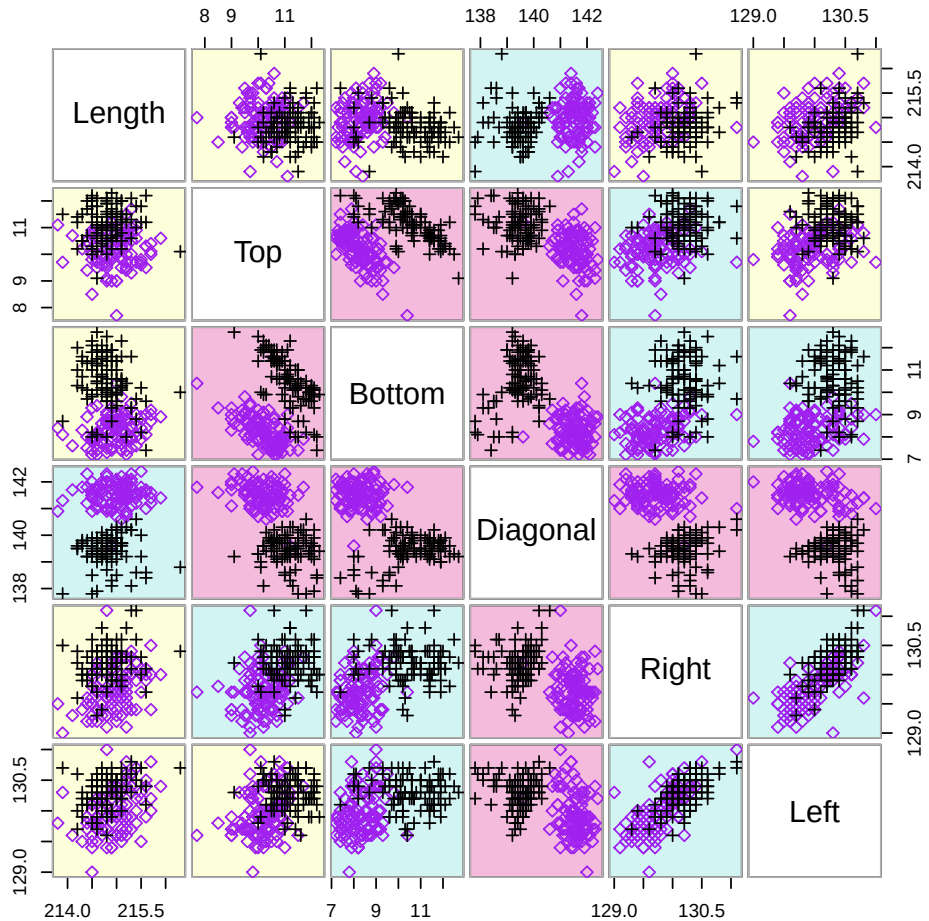
Swiss bank note data



Swiss bank note data

- The diagonal measurement seperates good from counterfit.
- Groups crossing in a parallel coordinate plot are indicative of negative correlation. eg Right thickness and length
- A lot of seperation may be in rotated dimensions, two ways to look

Swiss bank note data - splom



Lattice graphics

Lattice is an R implementation of the trellis framework developed by William Cleaveland and Rick Becker in the mid 90's at Bell Labs. Lattice is written by Deepayan Sarkar.

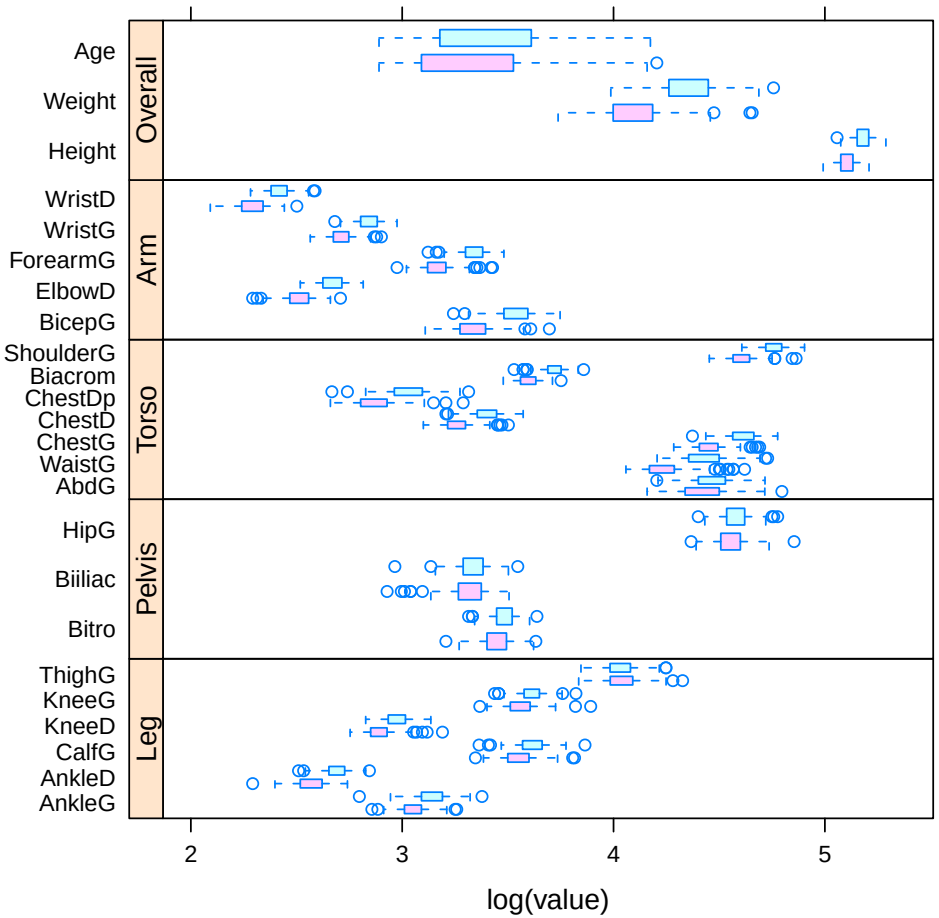
- Lattice is built on the grid graphics system, so commands in lattice have to be grid commands
- Panels are laid into rows and columns so that each panel shows a subset of the data.
- Panels can be created using different variables, or by conditioning.
- The panel displays are controled through panel functions.

The body data

this dataset contains 21 body dimension measurements as well as age, weight, height, and gender on 507 individuals.

- 247 men, 260 women
- Individuals in their twenties and thirties (A few older) all exercising regularly
- Citation: Heinz, G., Peterson, L.J., Johnson, R.W. and Kerk, C.J. (2003), "Exploring Relationships in Body Dimensions", Journal of Statistics Education.

Plotting body data

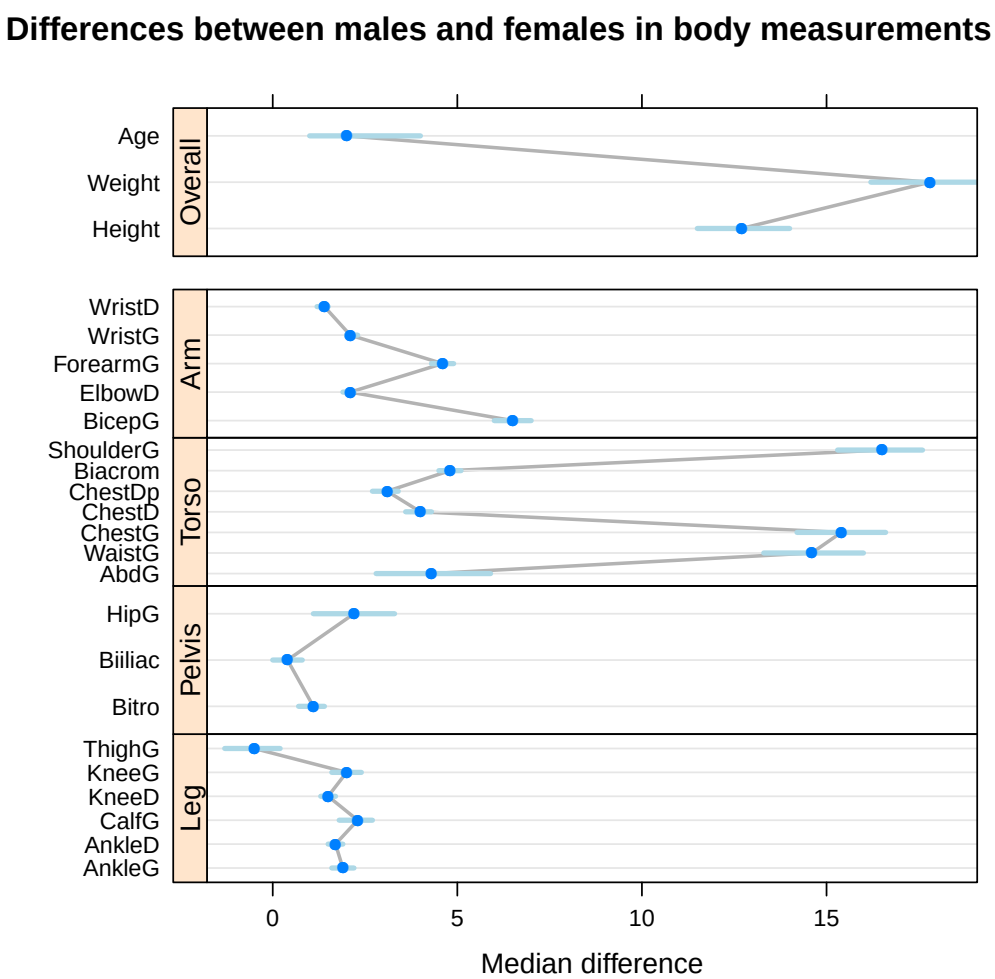


Perceptual sorting

It is worth noting that the figure was sorted and grouped

- grouped by position on the human body
- sorted, so that the order goes from hands->torso->hips->legs
- The plot could be further improved,
  - Differences are perceptually hard to distinguish across many comparisons, plotting the differences themselves is better.

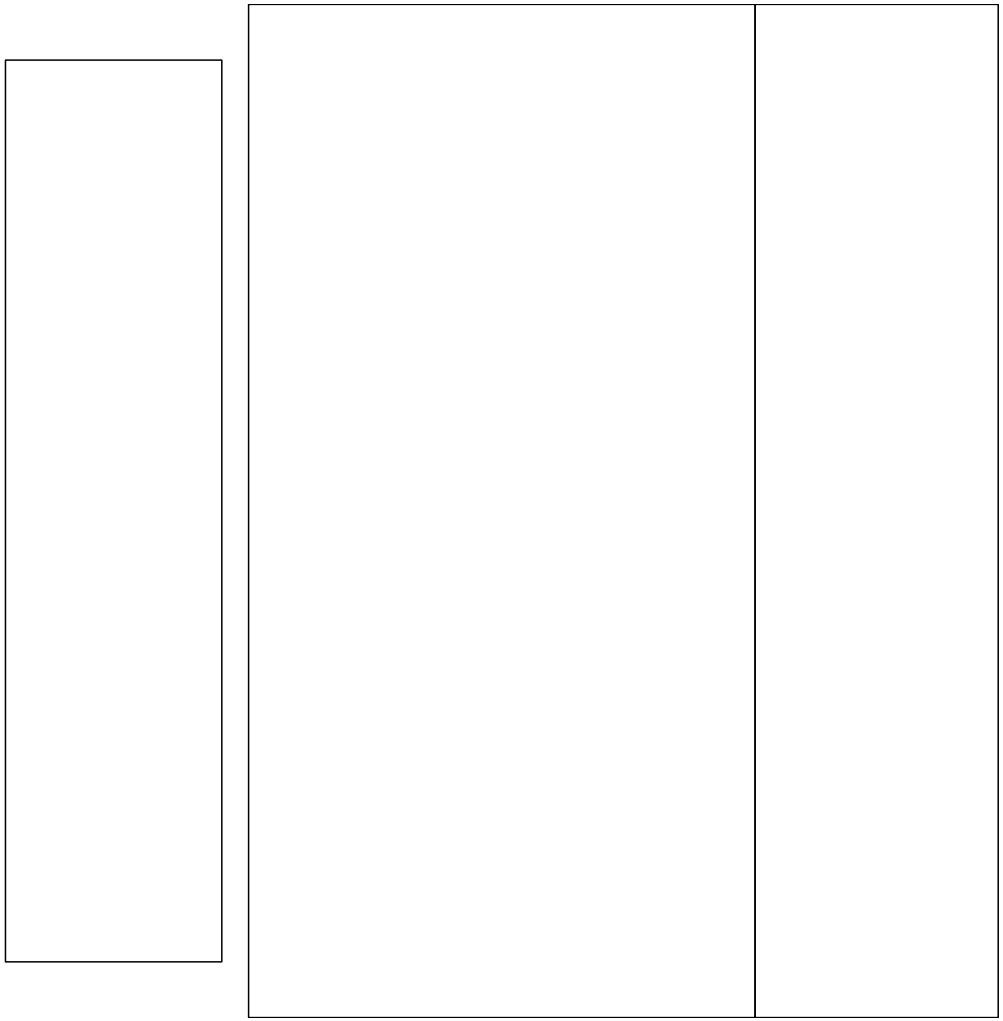
Plotting body data - 2



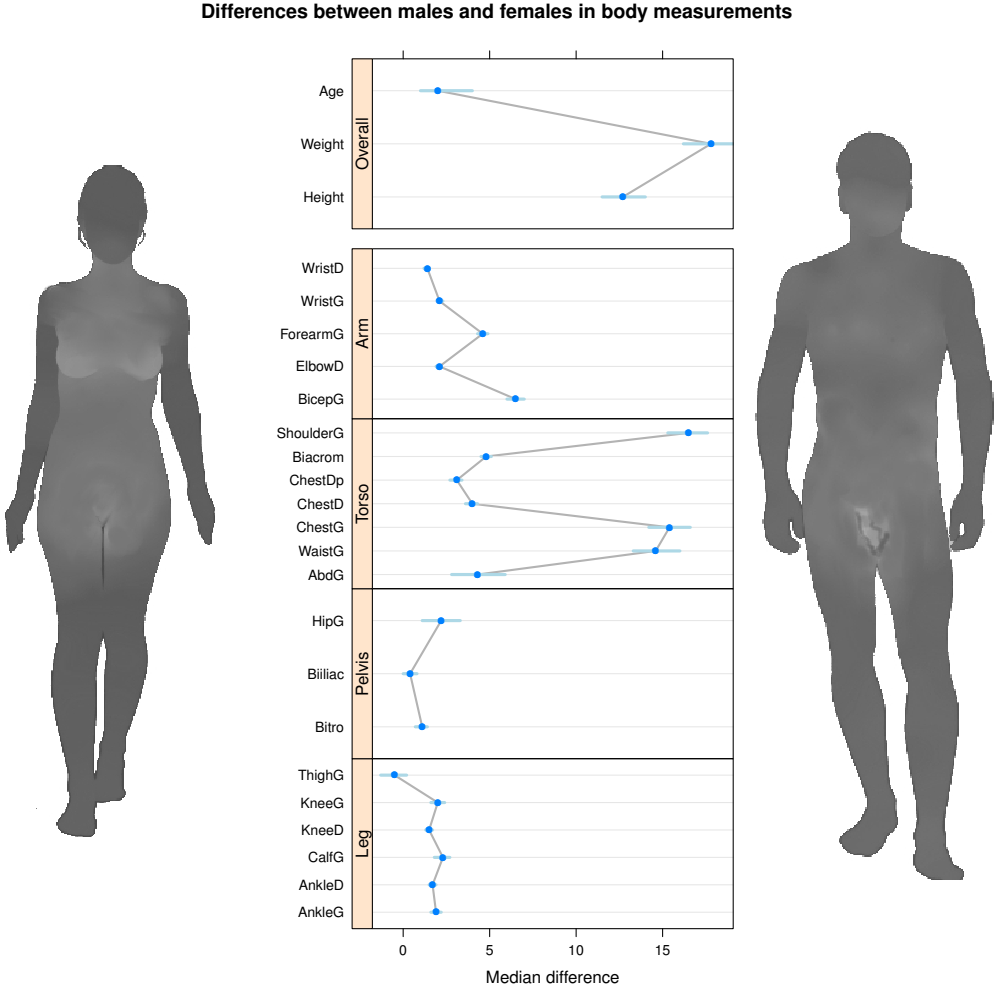
Enhancing further

- Since lattice lattice is based on grid, one should be able to plot into a viewport? Yes, exactly

Enhancing further



Adding eye candy



Overplotting and rendering time

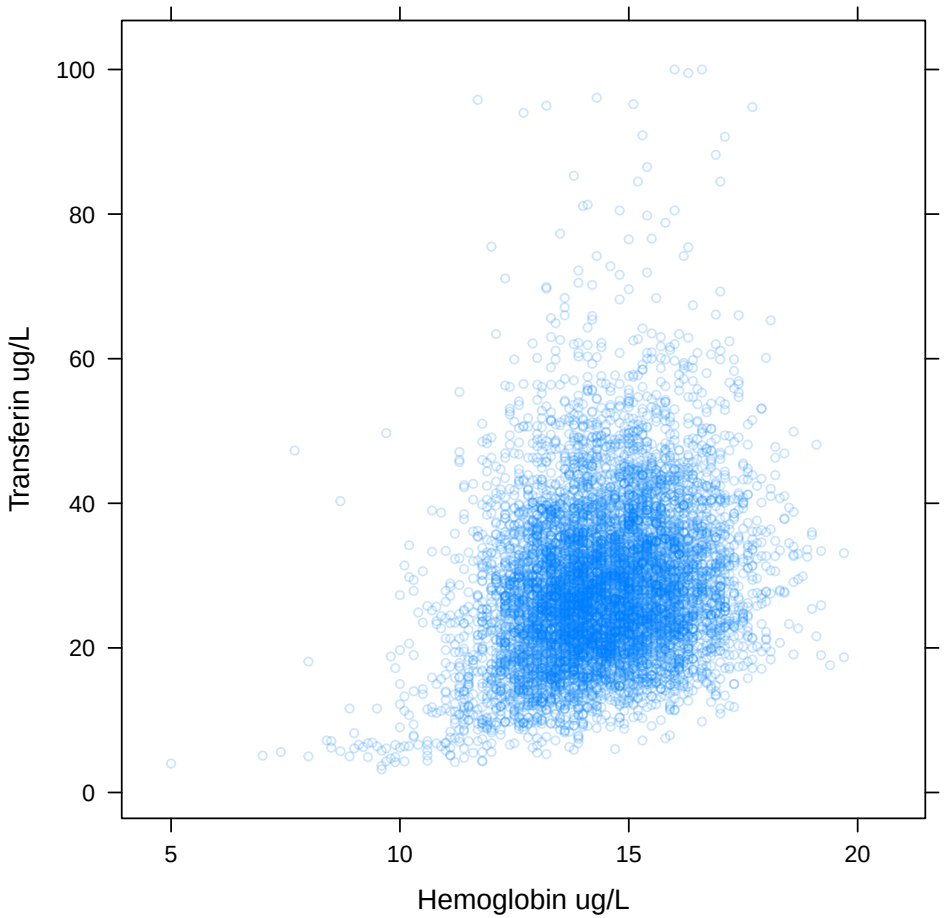
As we saw earlier, if the data gets too large the rendering time becomes extermely slow, Also

- Overplotting can become quites severe
- One solution is alpha blending, which cures overplotting, but rendering still slow
- Binning with judicious use of color can solve the problem

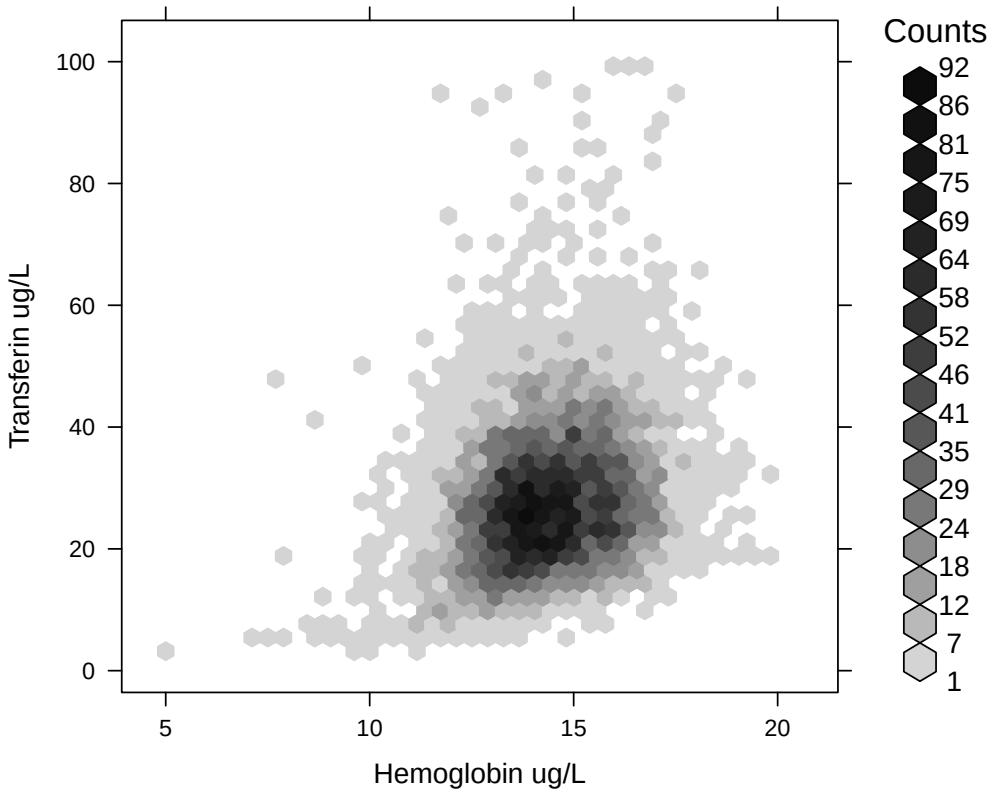
National Health and Nutrition Examination Survey, NHANES

- 15 variables (columns) on 9575 persons (rows)
- 9 continuous variables, 6 factors
- Not huge, but large data set.

Hexagon binning



Hexagon binning



Special prepanel and panel functions

- Hexagon plots need to be scaled properly so that aspect ratios do not distort the hexagons.
- Requires special prepanel functions to calculate the aspect ratio of each panel, so it is available to the panel function.
-

ggplot2

Created in the last few years by Hadley Wickham, `ggplot` is an R implementation of Leland Wilkonson's "Grammer of Graphics"

- Uses a very different object and computational model than other R graphics packages
  - Based on prototype objects, each operation modifies a single prototype instantation
  - all computations done in the environment of the proto object
- Very feature rich, easier to do many operations than lattice, but if the feature doesn't exaist can be harder.

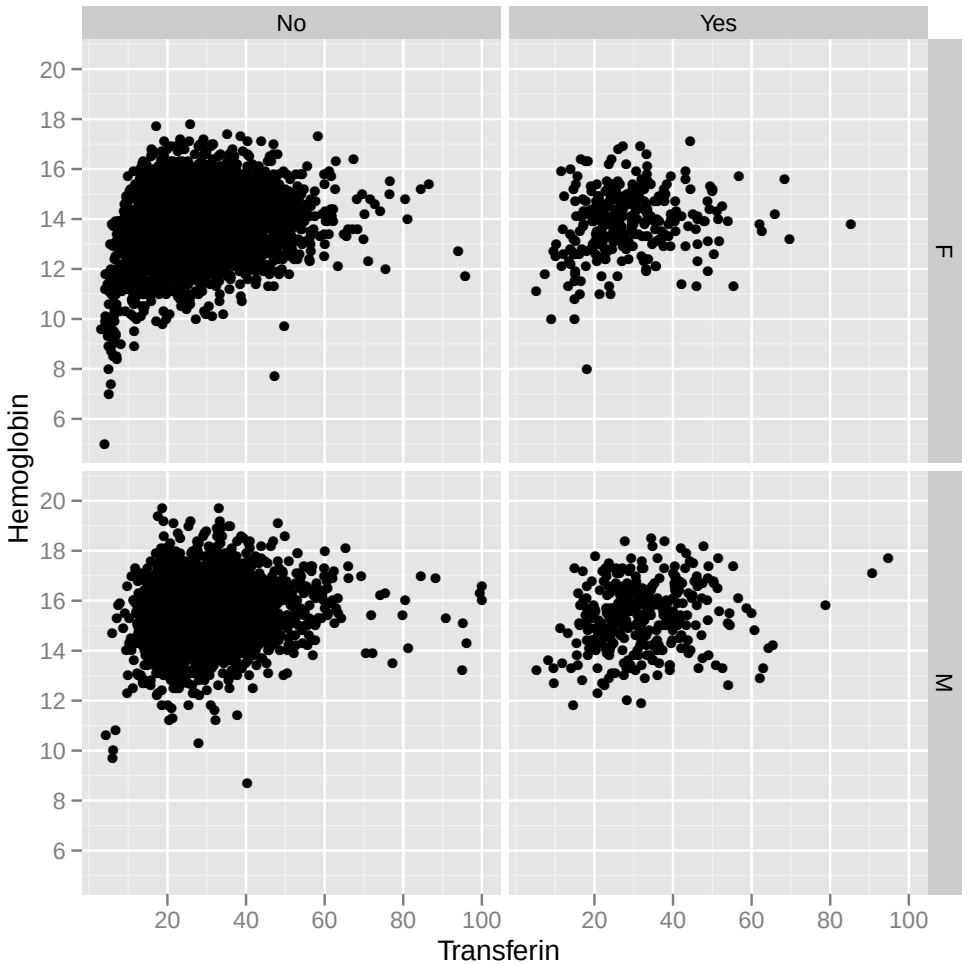
Two ways of creating plots

- `qplot`
  - Allows compact specification of plotting commands using a more "R" like syntax
  - Less control of the plot details
- `ggplot`
  - Very different syntax from usual R plotting
  - Much finer control of graphics output
  - Steeper learning curve.

Making a plot

```
> p <- qplot(Transferin, Hemoglobin, data = NHANES, facets
+     Cancer.Incidence)
```

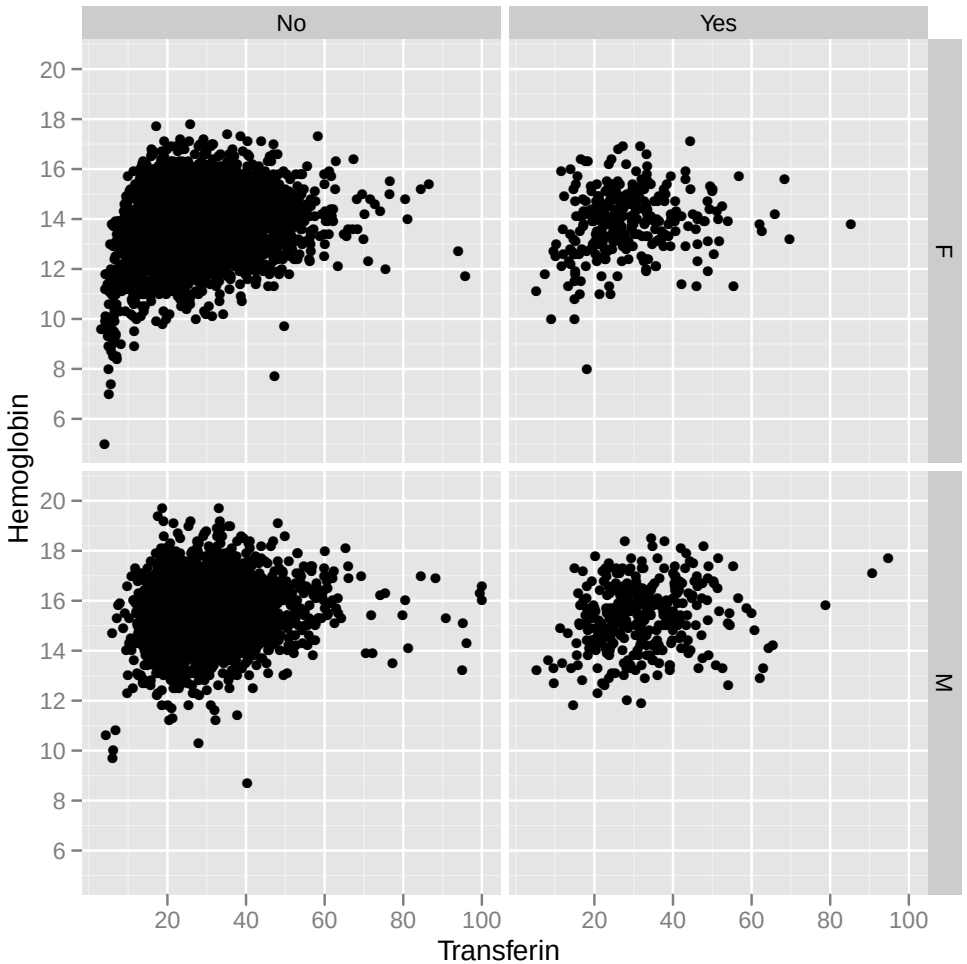
Making a plot



Making a plot

```
> p + geom_point(aes(colour = Cancer.Death))
```

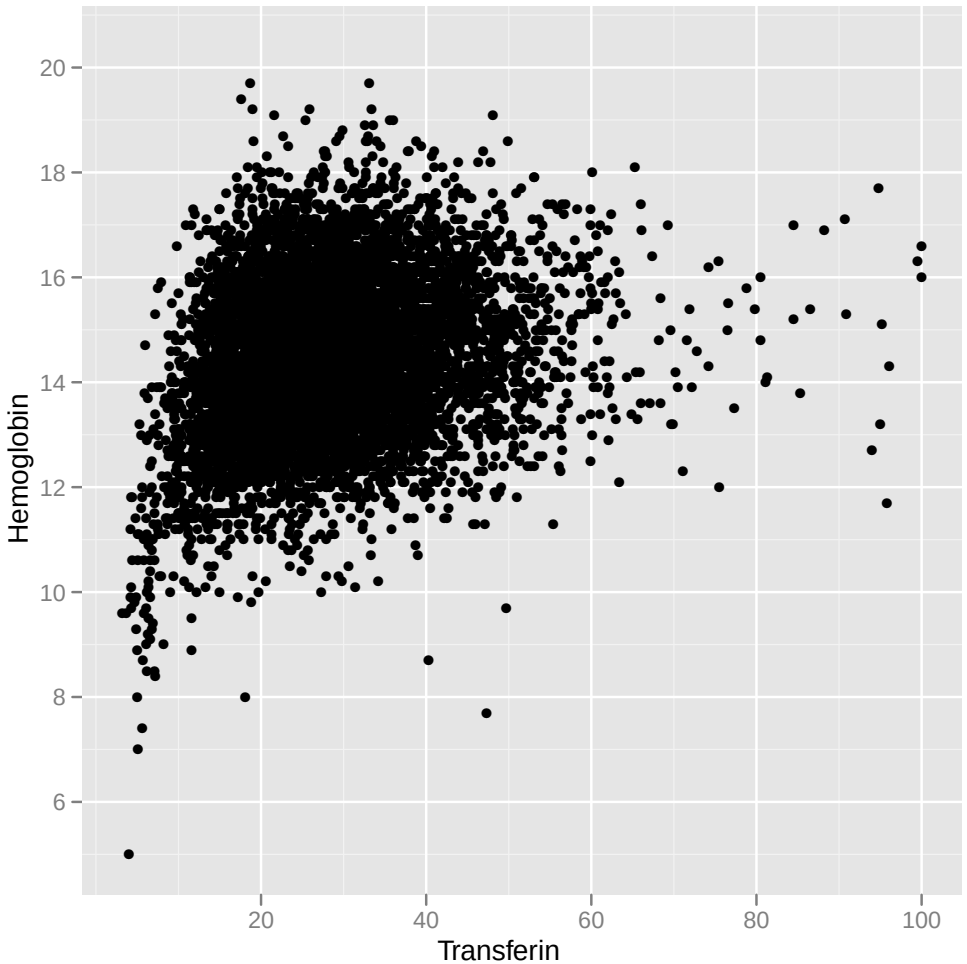
Making a plot



Doing it in steps

```
> p <- ggplot(NHANES, aes(x = Transferin, y = Hemoglobin))
> p <- p + geom_point()
```

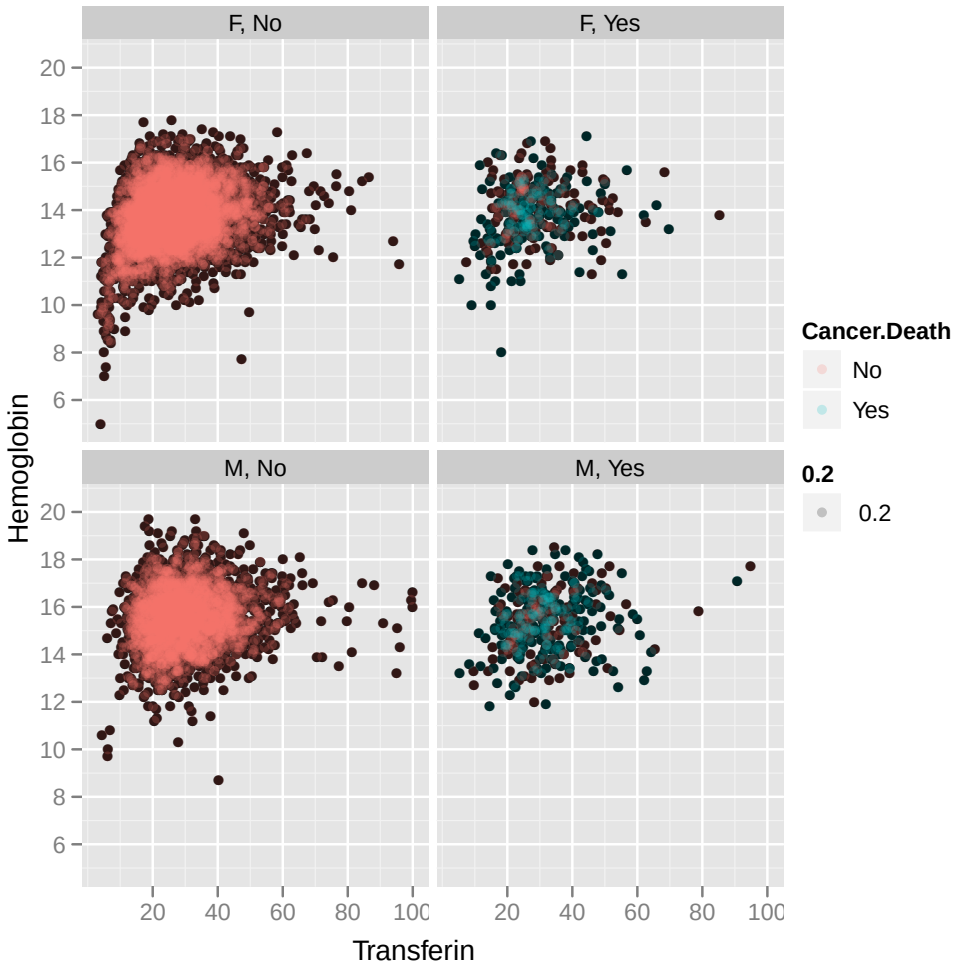
Doing it in steps



Doing it in steps

```
> p <- p + geom_point(aes(colour = Cancer.Death, alpha = 0.
+     facet_wrap(Sex ~ Cancer.Incidence)
```

Doing it in steps



Other systems

- `rgl` A package with it's own viewer for interactive 3d viewing.
- `iplots` Interactive graphics using java, based on Mondrian
- `rggobi` Interactive graphics using GTK
- Many others in the pipeline, will know more after Use!R and DSC.

End

- I make no claim that any system is better than another, they are all fun.
- I am only scratching the surface, there is so much more, ...
- If people are interested I can do something next year on interactive graphics (iplots, rggobi, ...)

Bibliography

- Hurley, C. 2004. Clustering visualizations of multidimensional data. JCGS 13(4): 788-806
- Murrell, P. 2005. R Graphics, CRC Press
- Sarkar, D 2007. Lattice: Multivariate Data Visualization with R, Springer, NY
- Wickham, H. (In Press). ggplot: Elegant graphics for data analysis. Springer, NY