Correlation Plots Note

Mick McQuaid

```
library(knitr)
opts_chunk$set(tidy = FALSE)
```

Here are my brief notes about correlation plots, graphical displays of correlation matrices. You can find a detailed introduction to the `corrplot` package at `https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html`

Correlation matrices make good summaries of pairwise comparisons but can be difficult to read, often compressing many numbers into a small square. Correlation plots make it easier to read correlations at a glance.

Following is my favorite way to use the `corrplot` package, with the upper diagonal containing a display of ellipses and the lower diagonal containing the Pearson sample correlation coefficient,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

This coefficient ranges from 1, meaning perfectly positively correlated to -1, meaning perfectly negatively correlated. What you see below is the corrplot for `mtcars`, a data set including information about a sample of cars marketed in the year 1974.

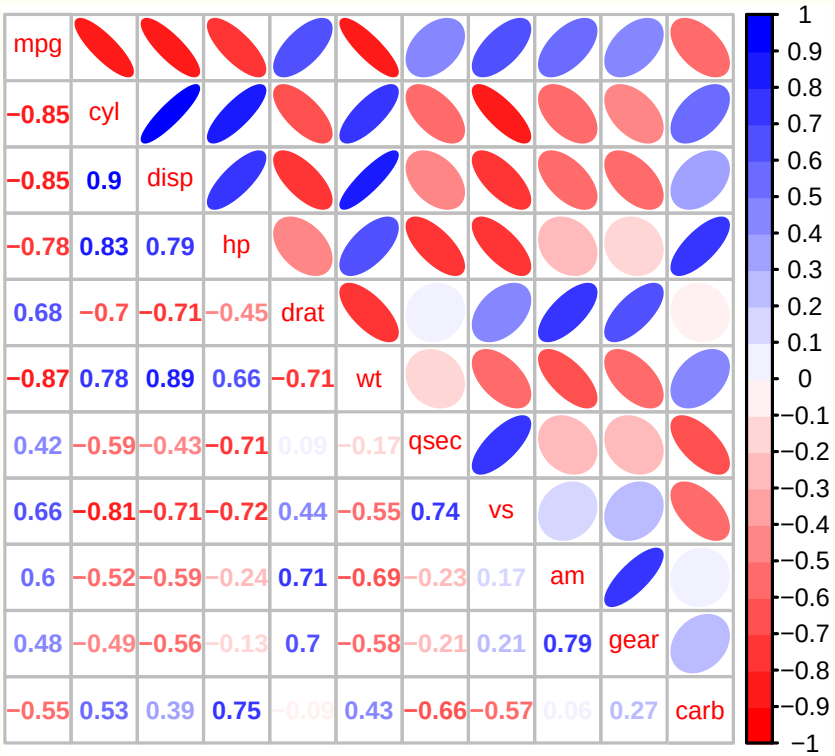
The scale of colors for the entries ranges from dark blue for perfectly correlated, to brick red for perfectly negatively correlated. The shape and direction of the ellipses varies similarly. For example, the most highly correlated variables are `disp` and `cyl` at 0.9 and their ellipse is narrow, dark blue, and points to the upper right and lower left. The least correlated are `am` and `carb` at 0.06 and their ellipse is nearly a circle and is nearly white.

The correlation matrix (and hence the plot) is symmetric along the main diagonal: all the information in the upper diagonal would be repeated in the lower diagonal so it makes sense to use the cells differently, in this case displaying the actual numbers in the lower diagonal and the ellipses that represent the numbers in the upper diagonal. The `corrplots` package has many different ways to display the correlation plot and many different color schemes as well.

You can not run the following code directly on a data set like the NYC

311 data because it is not numeric. But you can create numeric tables on which the following code would work by, for instance, creating incidence matrices of the frequency with which classes of complaints are made in each borough. Then you could see which boroughs are more similar or more different or which complaints are more similar or more different.

```
library(corrplot)
m <- cor(mtcars)
col3 <- colorRampPalette(c("red", "white", "blue"))
corrplot.mixed(m,
            upper="ellipse",
            upper.col=col3(20),
            lower.col=col3(20),
            number.cex=0.5,
            tl.cex = 0.5,
            cl.cex=0.5)
```



`number.cex` controls the size of the correlation coefficients

`tl.cex` controls the size of the labels on the diagonal

`cl.cex` controls the size of the numbers on the scale (−1 to 1)