Exploratory Analysis

IS 665 Data Mining, Data Warehousing and Visualization

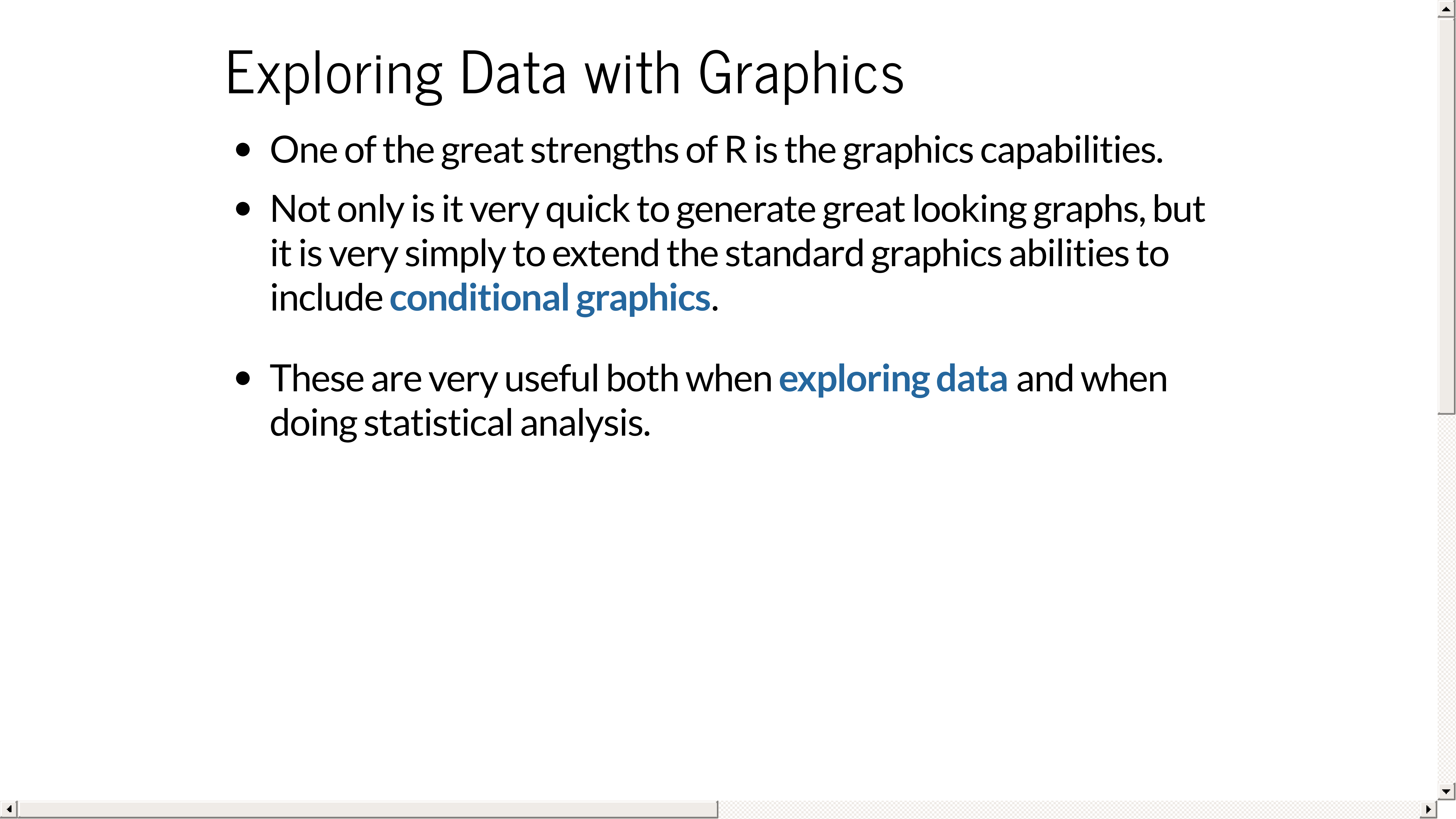
Exploratory Data Analysis

The iterative process of:

- Generating questions about your data
- Searching for answers by visualing and transforming your data
- Refining your questions based on what you learned.

Hypothesis Testing vs Exploratory Analysis

- Hypothesis testing start with a conjecture about what to expect to see from the data (hypothesis)
- Exploratory Analysis start without any *a priori* hypotheses.
- Usually done with graphs and by looking at simple relationships between variables



Exploring Data with Graphics

- One of the great strengths of R is the graphics capabilities.
- Not only is it very quick to generate great looking graphs, but it is very simply to extend the standard graphics abilities to include **conditional graphics**.
- These are very useful both when **exploring data** and when doing statistical analysis.

Graphical Environments

- **Base package** provides the simplest graphs: easy to remember, provides low level of analysis.
  plot(), hist()…
- **Lattice** is more options to create higher level of analysis.
  - syntax is similar to base functions
  - visual aspects (color, font etc) are harder to its alternatives (i.e. ggplot)
  - Good refrence: http://www.statmethods.net/advgraphs/trellis.html
- **Ggplot** is also good for higher level of analysis.
  - very detailed and well-thought-out visual functions
  - syntax is harder to learn (but not too hard to remember once learned.)
  - Refernce: http://docs.ggplot2.org/current/

Base Graphics

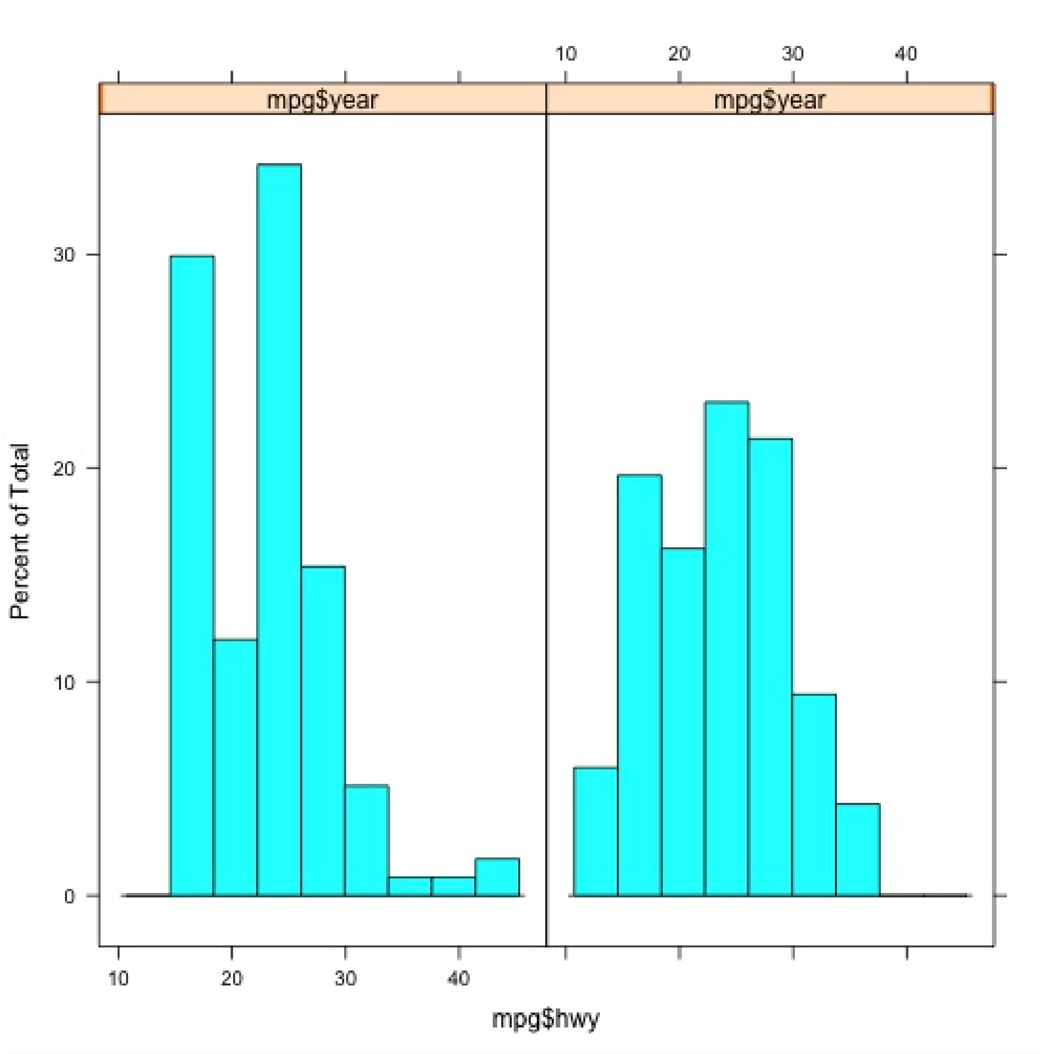
- plot: generic x-y plotting
- barplot: bar plots
- boxplot: box-and-whisker plot
- hist: histograms
- pie: pie charts
- dotchart: cleveland dot plots
- image, heatmap, contour, persp: functions to generate image-like plots
- qqnorm, qqline, qqplot: distribution comparison plots
- pairs, coplot: display of multivariant data

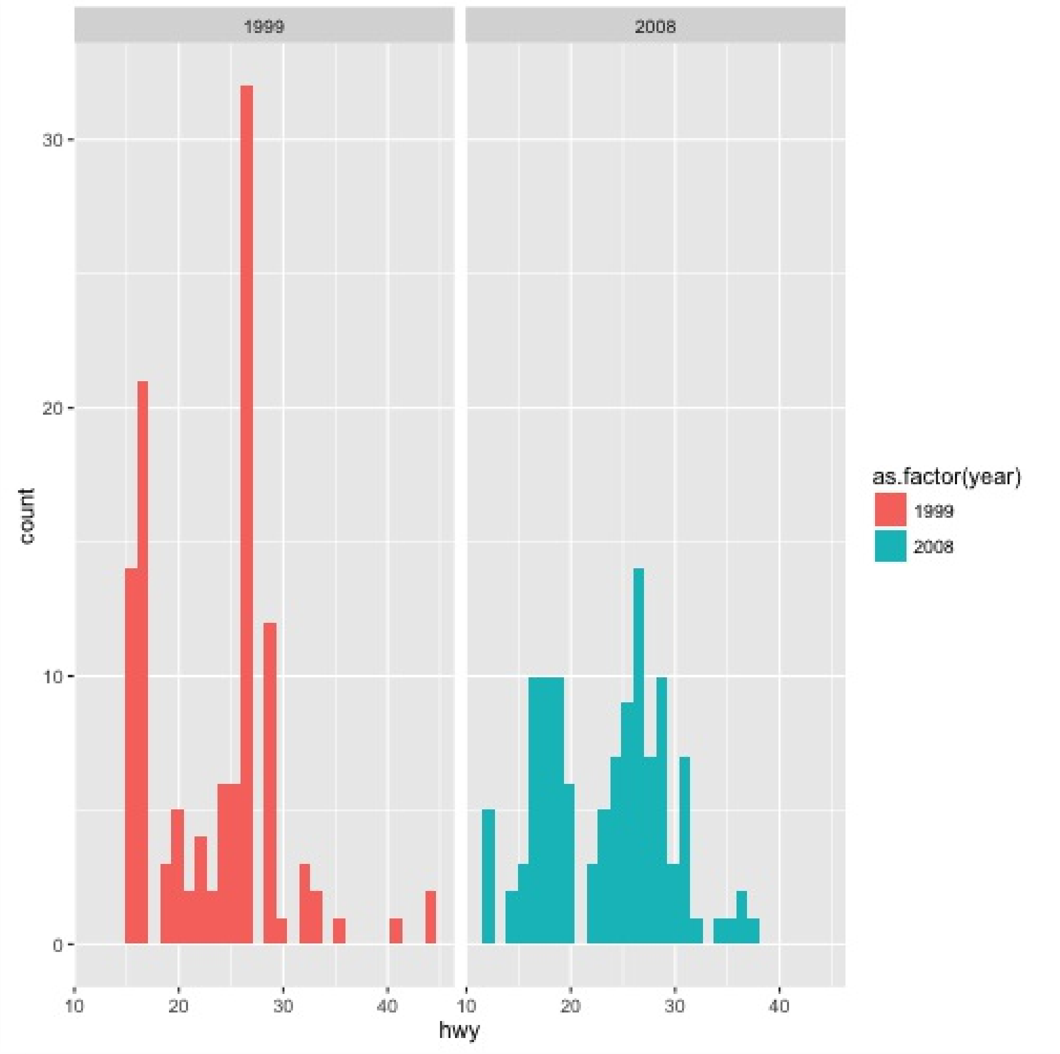
Lattice vs GGplot

Jury is still out on which is better

```
# install.packages('lattice') #if not installed already
require(lattice)
histogram(~mpg$hwy | mpg$year)
```

```
ggplot(mpg) + geom_histogram(aes(x = hwy, fill =
as.factor(year))) +
    facet_grid(~year)
```

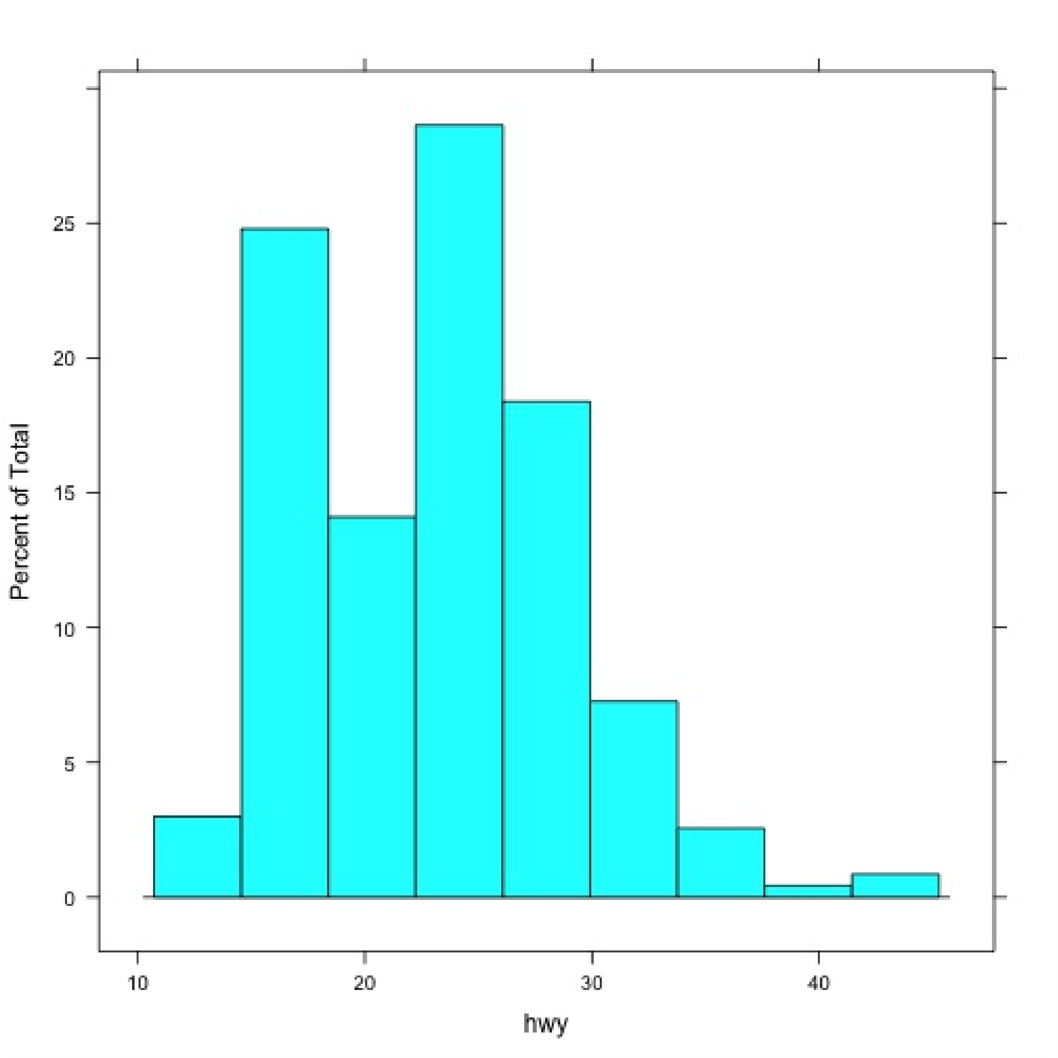


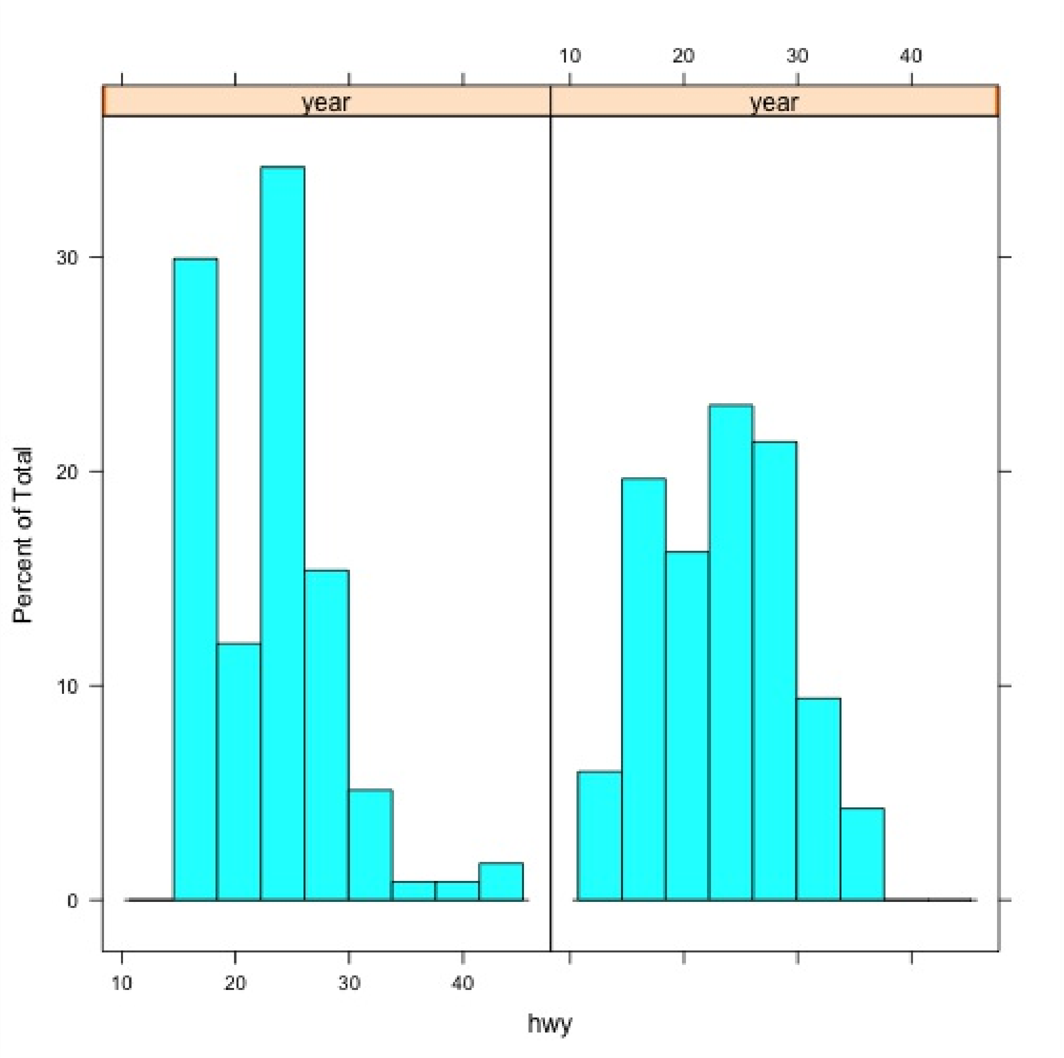


Histograms

```
# histograms
histogram(~hwy, mpg)
```

```
# histograms
histogram(~hwy | year, mpg)
```

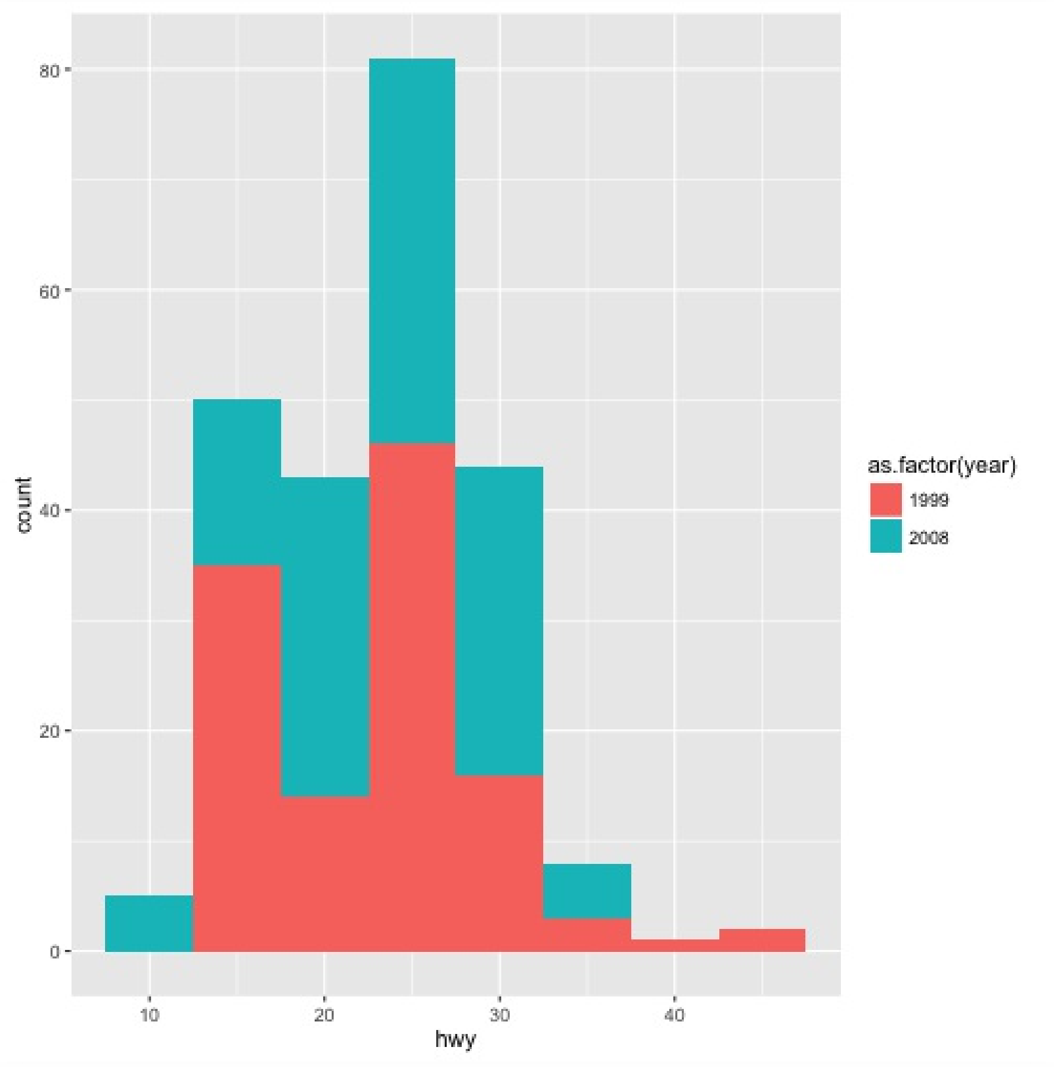


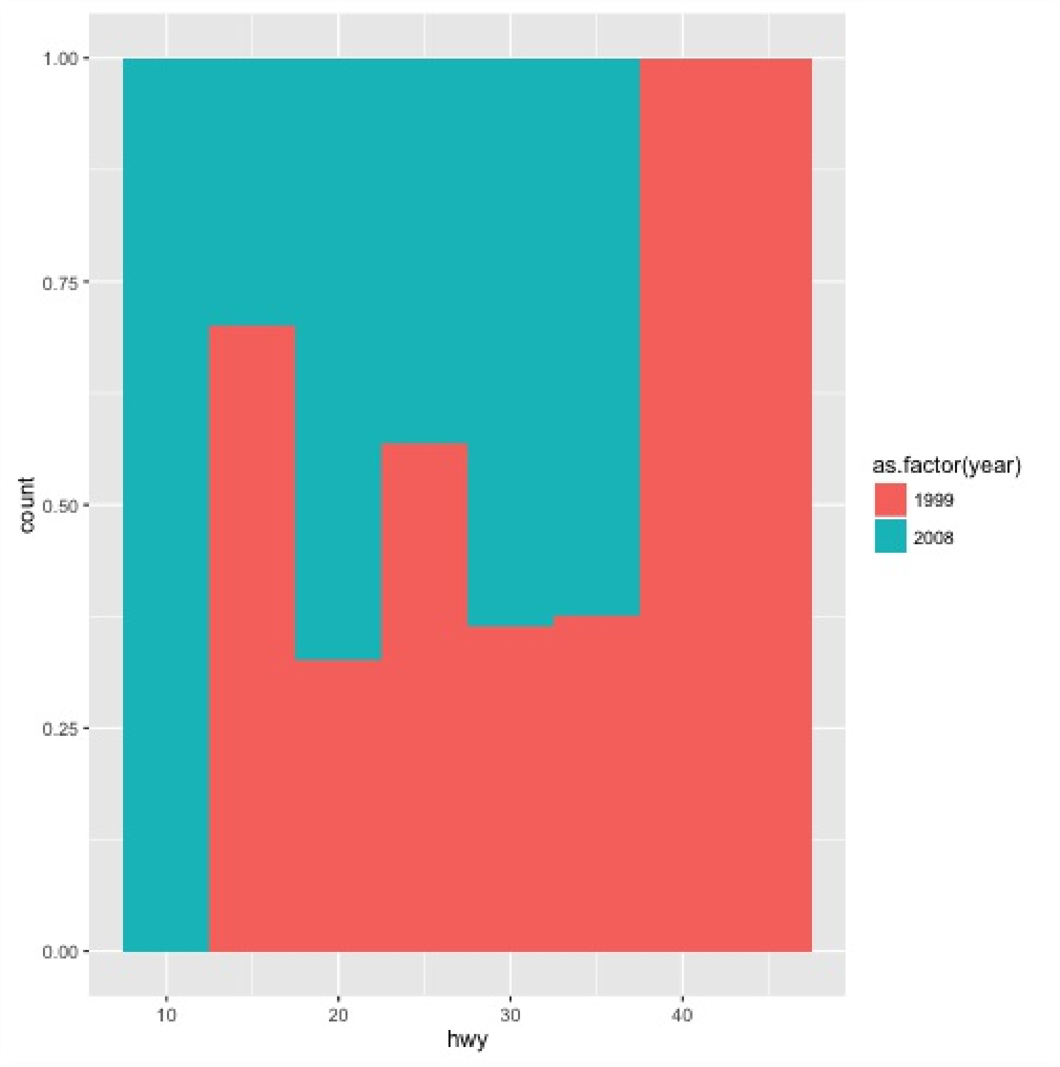


with ggplot

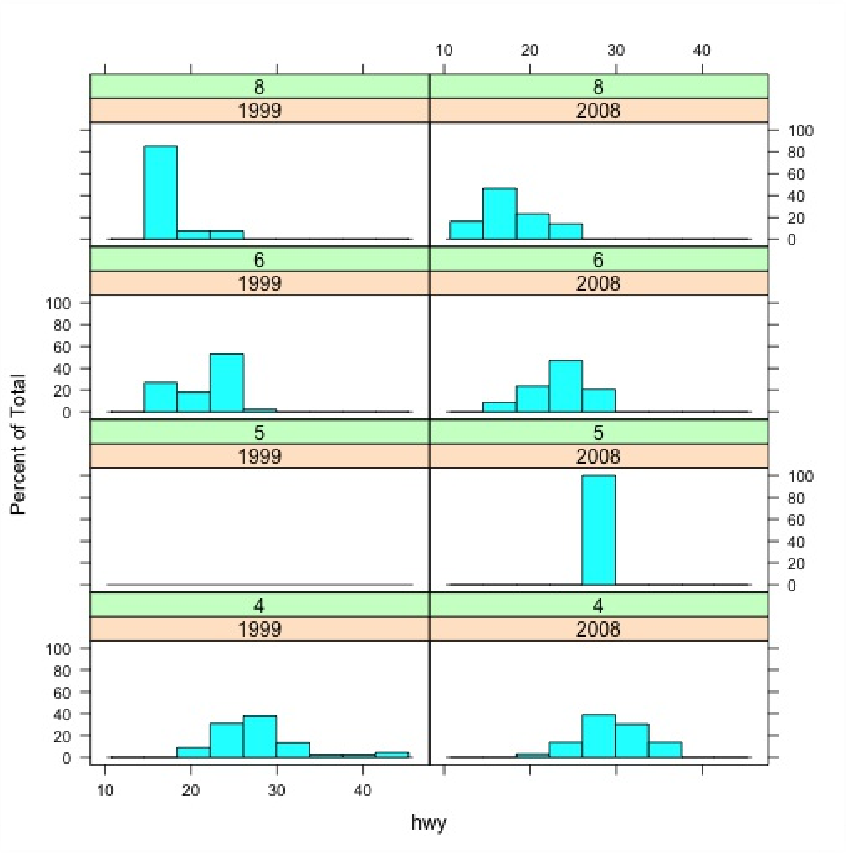
```
ggplot() + geom_histogram(data = mpg, aes(x =
hwy, fill = as.factor(year)),
    binwidth = 5)
```

```
ggplot() + geom_histogram(data = mpg, aes(x =
hwy, fill = as.factor(year),
    ), binwidth = 5, position = "fill")
```



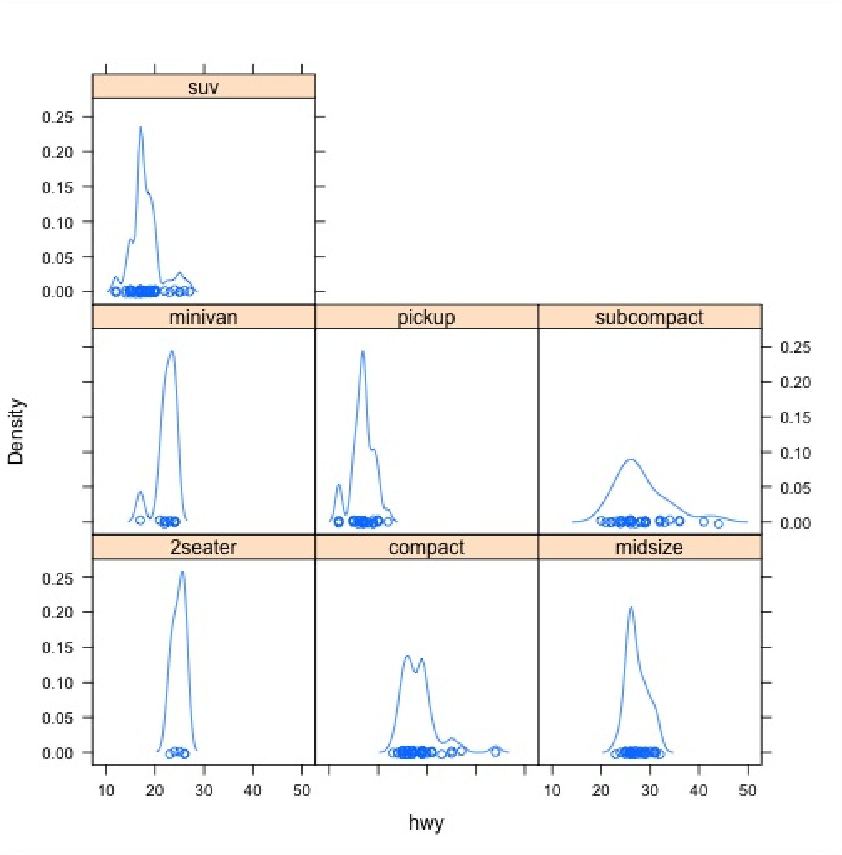


```
# histograms
histogram(~hwy | as.factor(year) + as.factor(cyl), mpg)
```

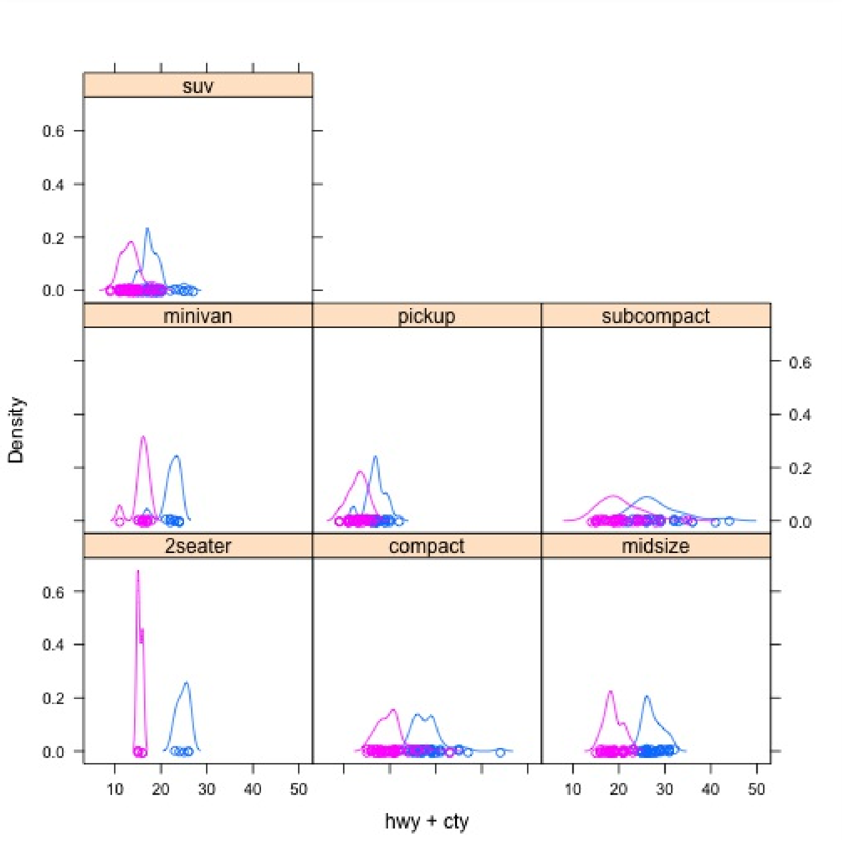


Density plots

```
densityplot(~hwy | class, mpg)
```



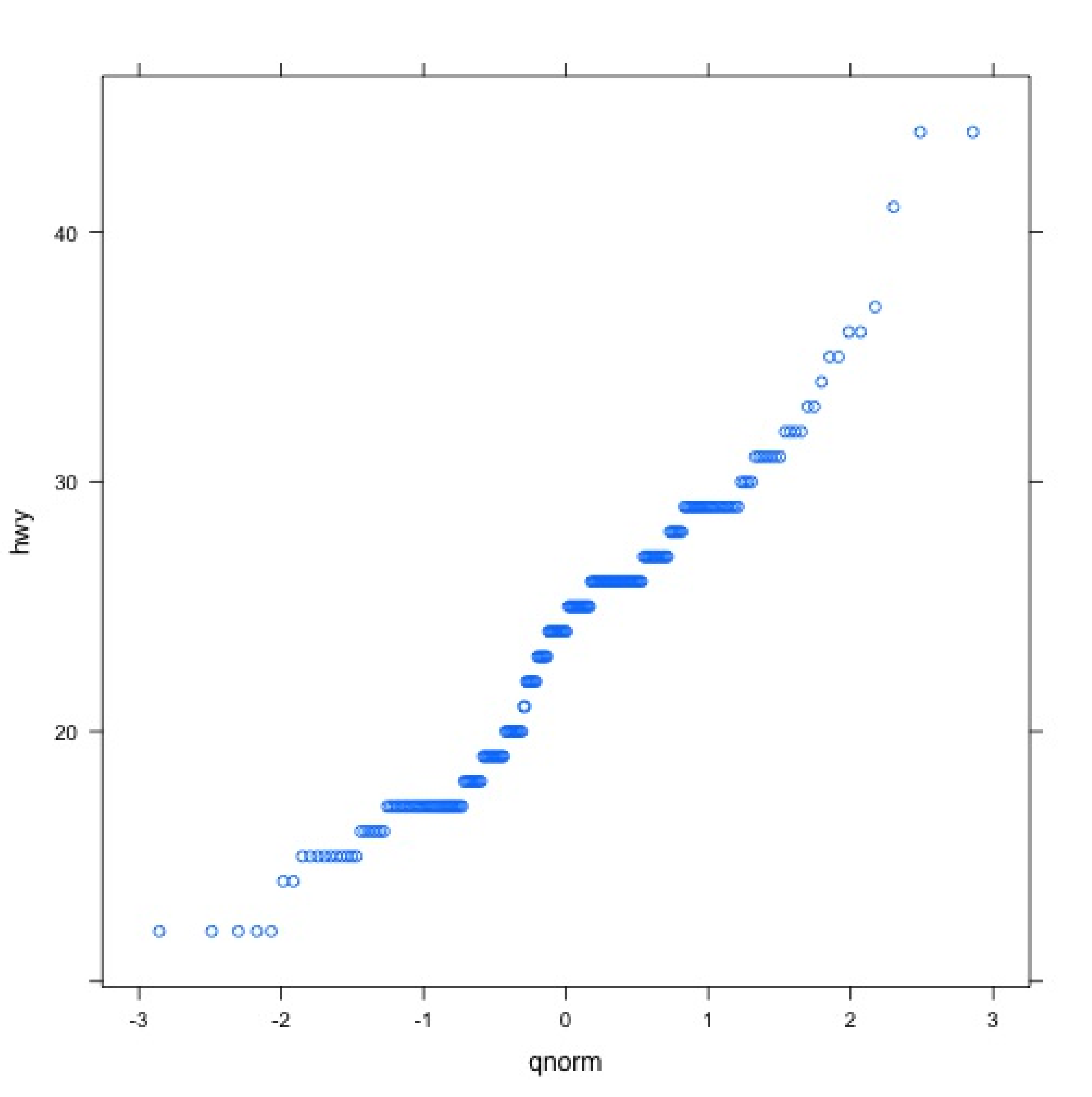
```
densityplot(~hwy + cty | class, mpg)
```

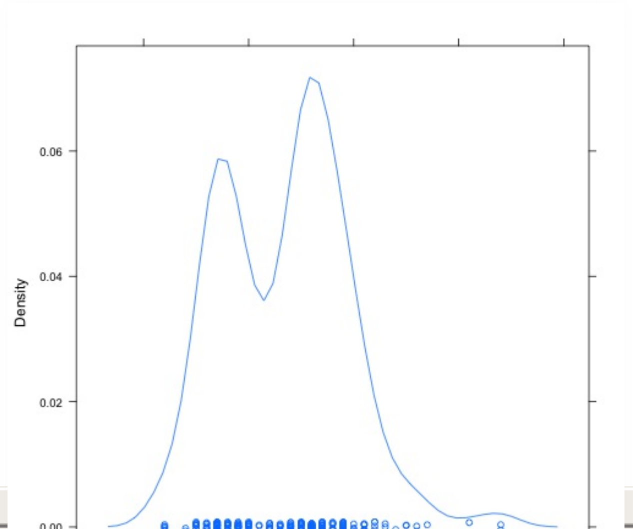


Q-Q Plots

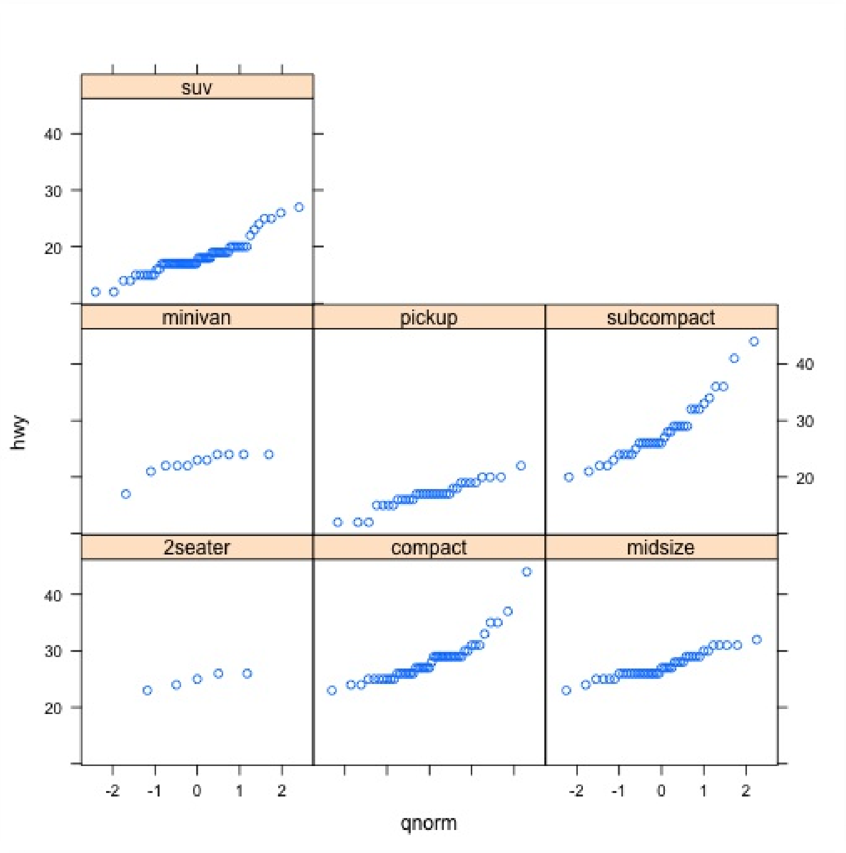
- To check if distributional assumptions are accurate.
- If points follow a straight line than assumptions are valid.

```
qqmath(~hwy, mpg)
```





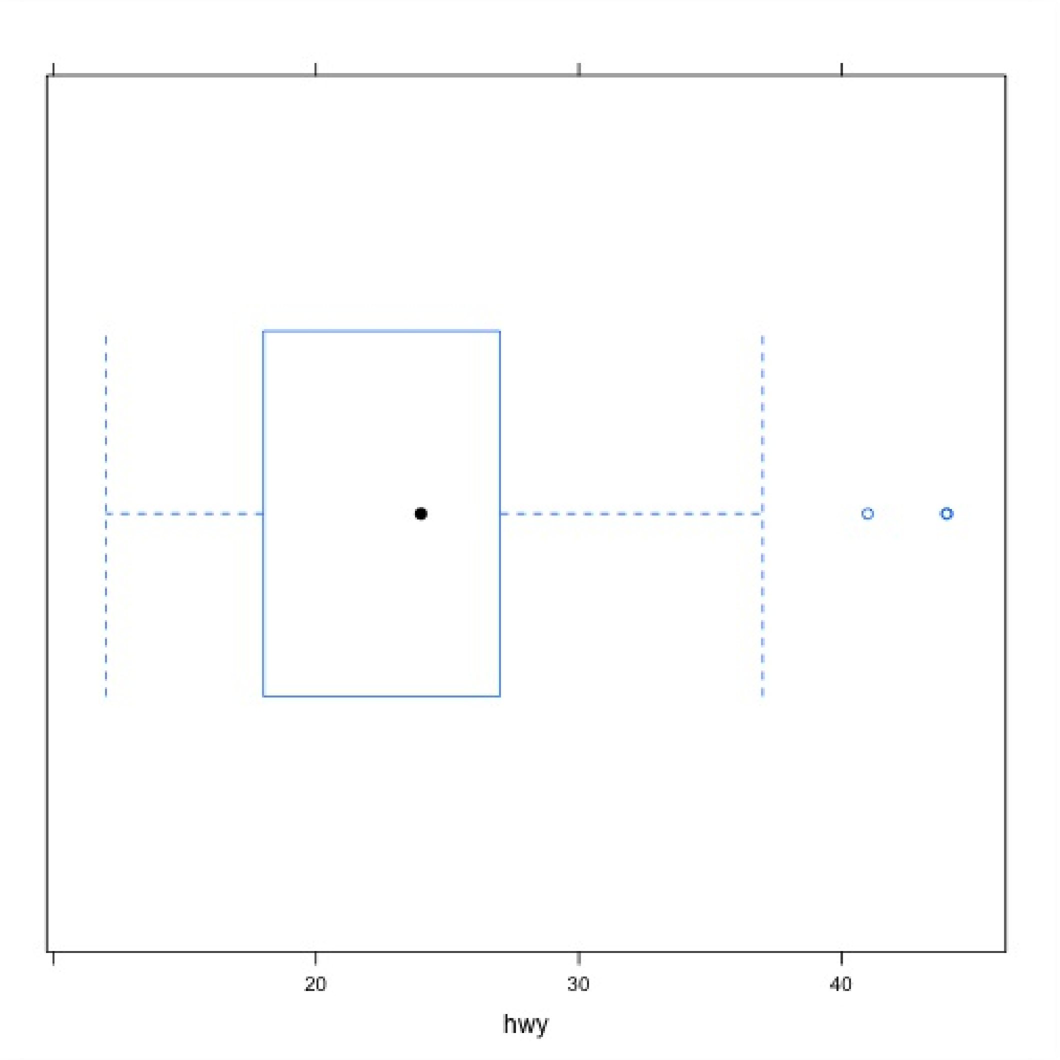
```
# conditional plot
qqmath(~hwy | class, mpg)
```



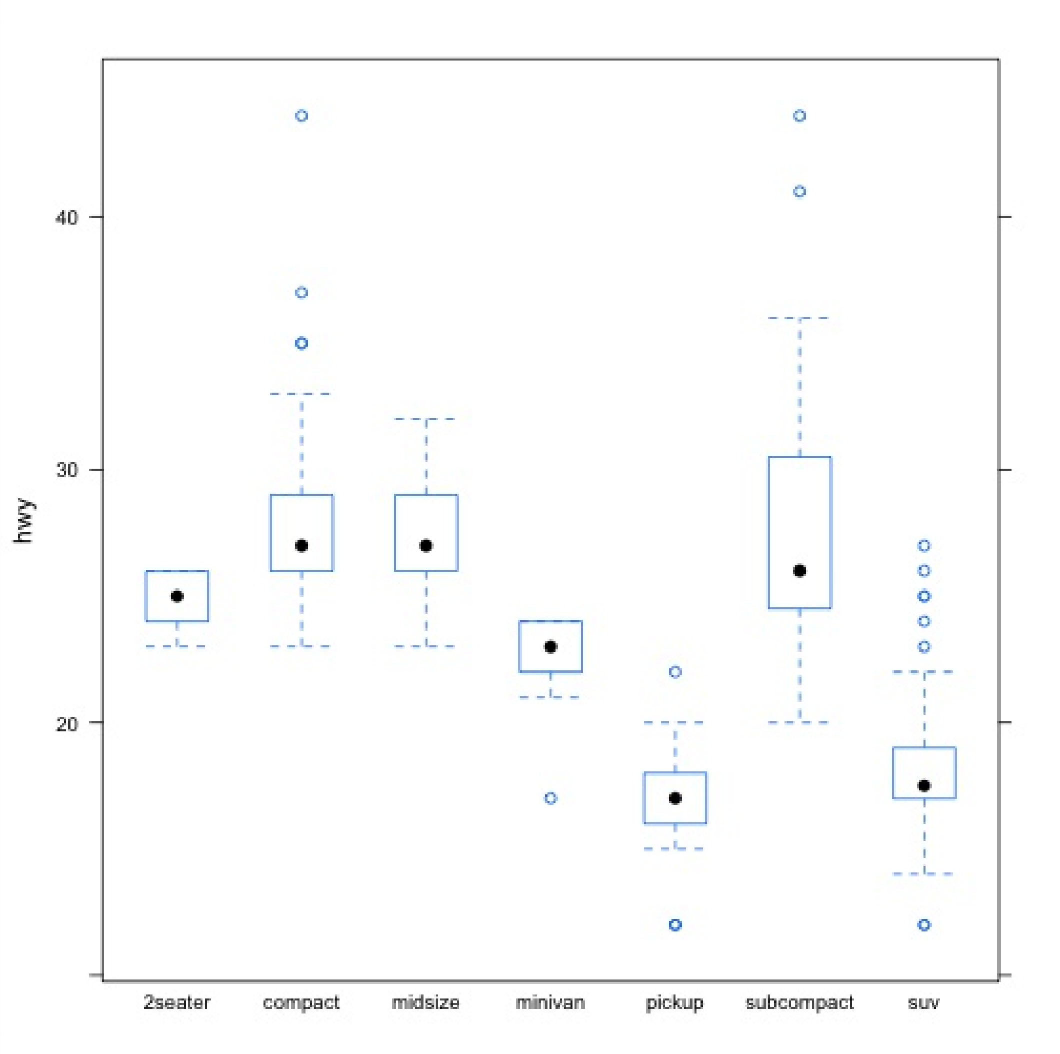
Box plots

A.k.a. Box and whiskers plots. Hence the command bwplot()

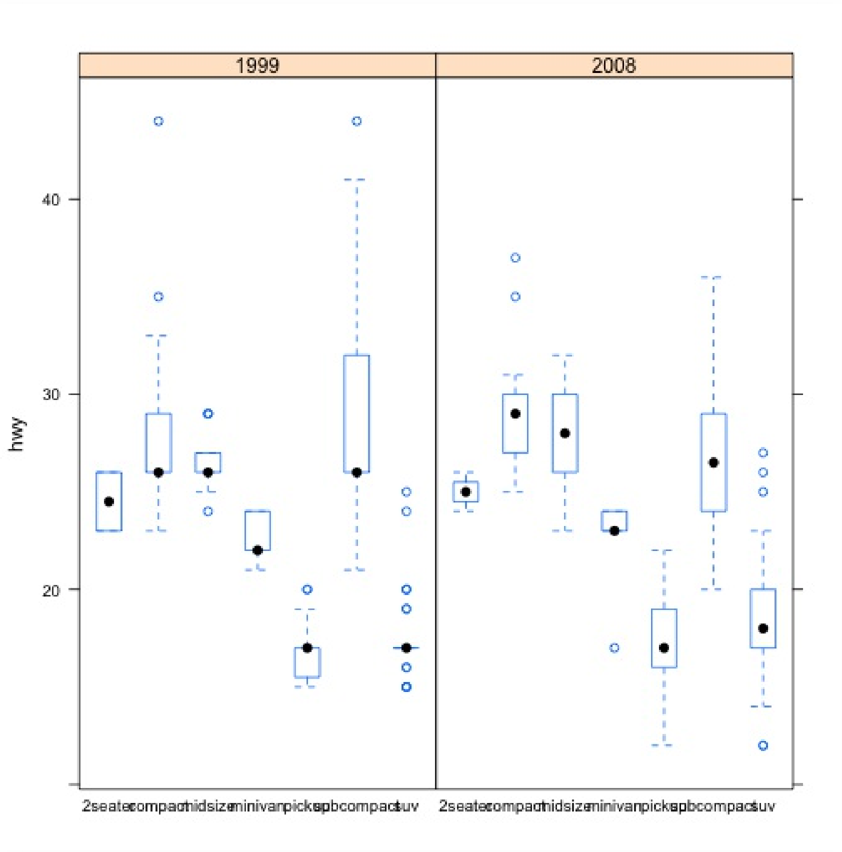
```
bwplot(~hwy, mpg)
```



```
# conditional
bwplot(hwy ~ class, mpg)
```



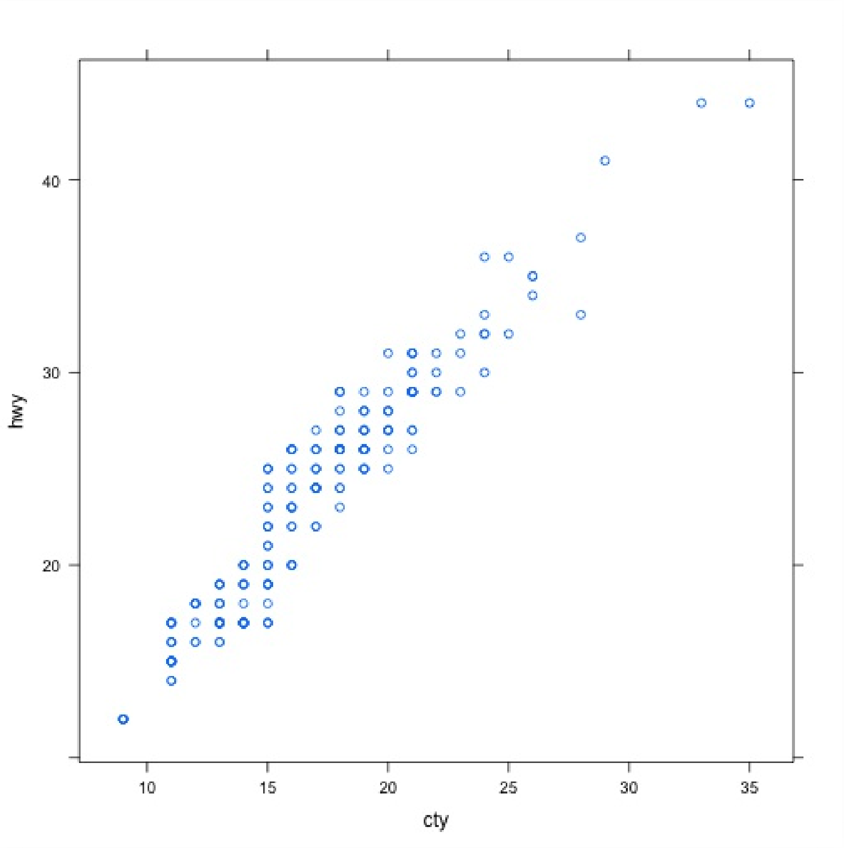
```r
# conditional
bwplot(hwy ~ class | as.factor(year), mpg)
```



Multivariate Plots

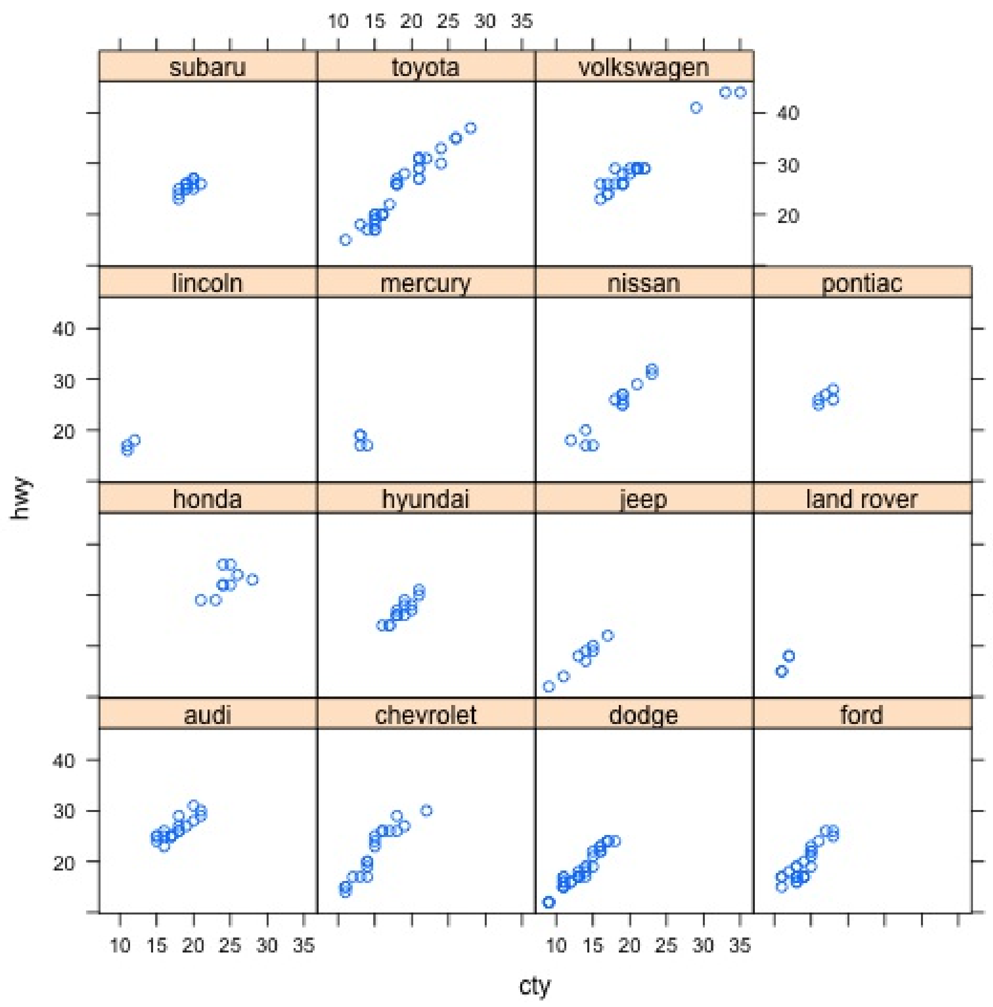
Scatter plots

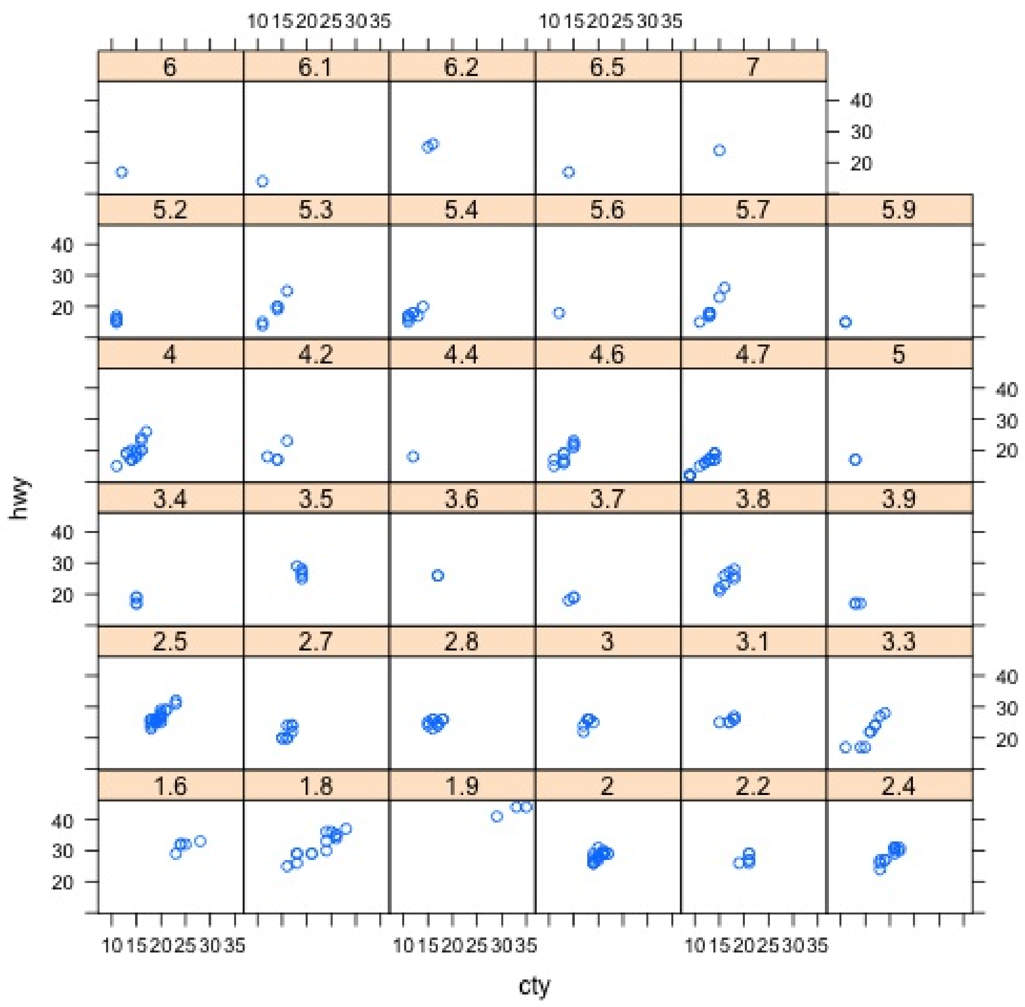
```
xyplot(hwy ~ cty, mpg)
```



```
xyplot(hwy ~ cty | manufacturer, mpg)
```

```
xyplot(hwy ~ cty | as.factor(displ), mpg)
```





Principles of Analytic Graphs

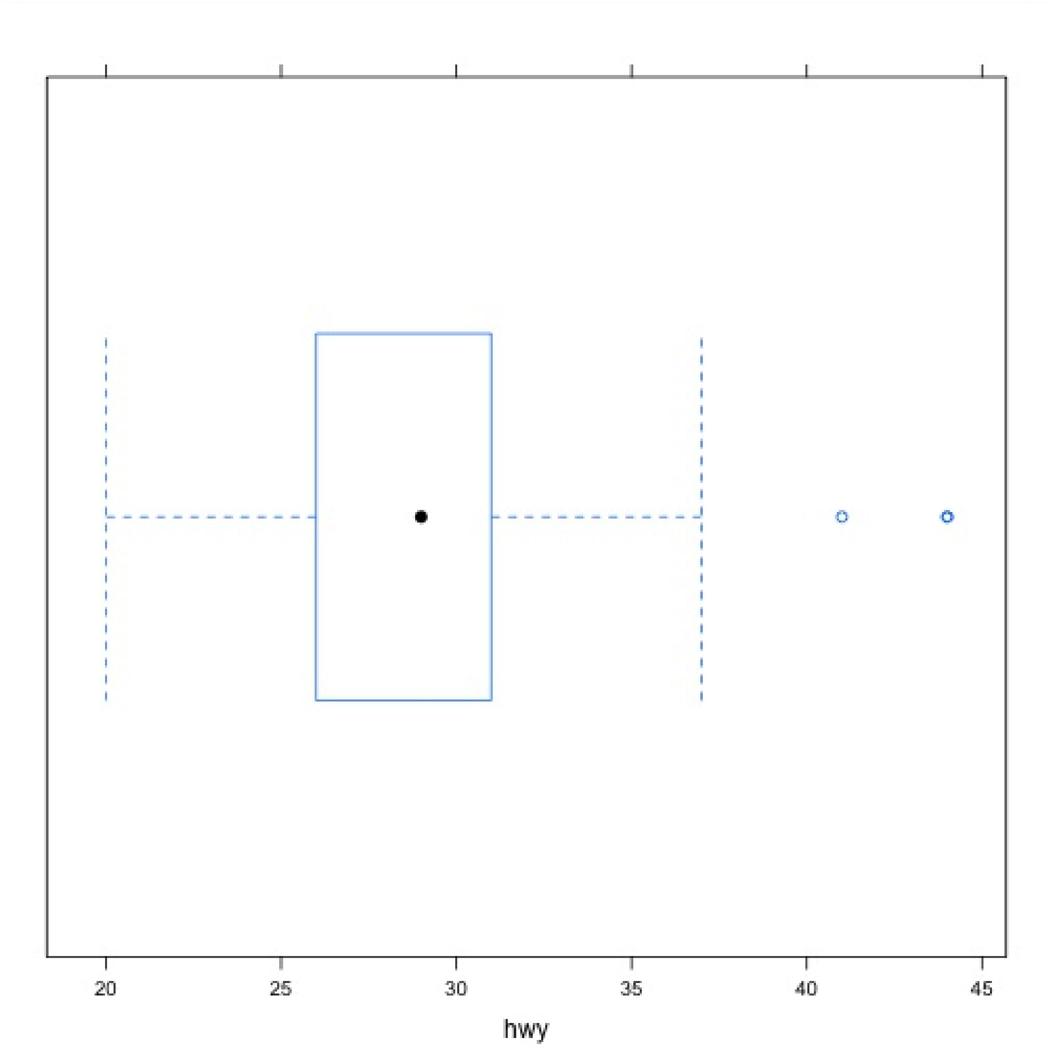
- Show comparisons
- Show causality, mechanism, explanation, systematic structure
- Show multivariate data
- Integrate evidence
- Describe and document the evidence
- Content, content, content

Derived from Edward Tufte’s wonderful book Beautiful Evidence via Exploratory Data Analysis with R by Roger D. Peng

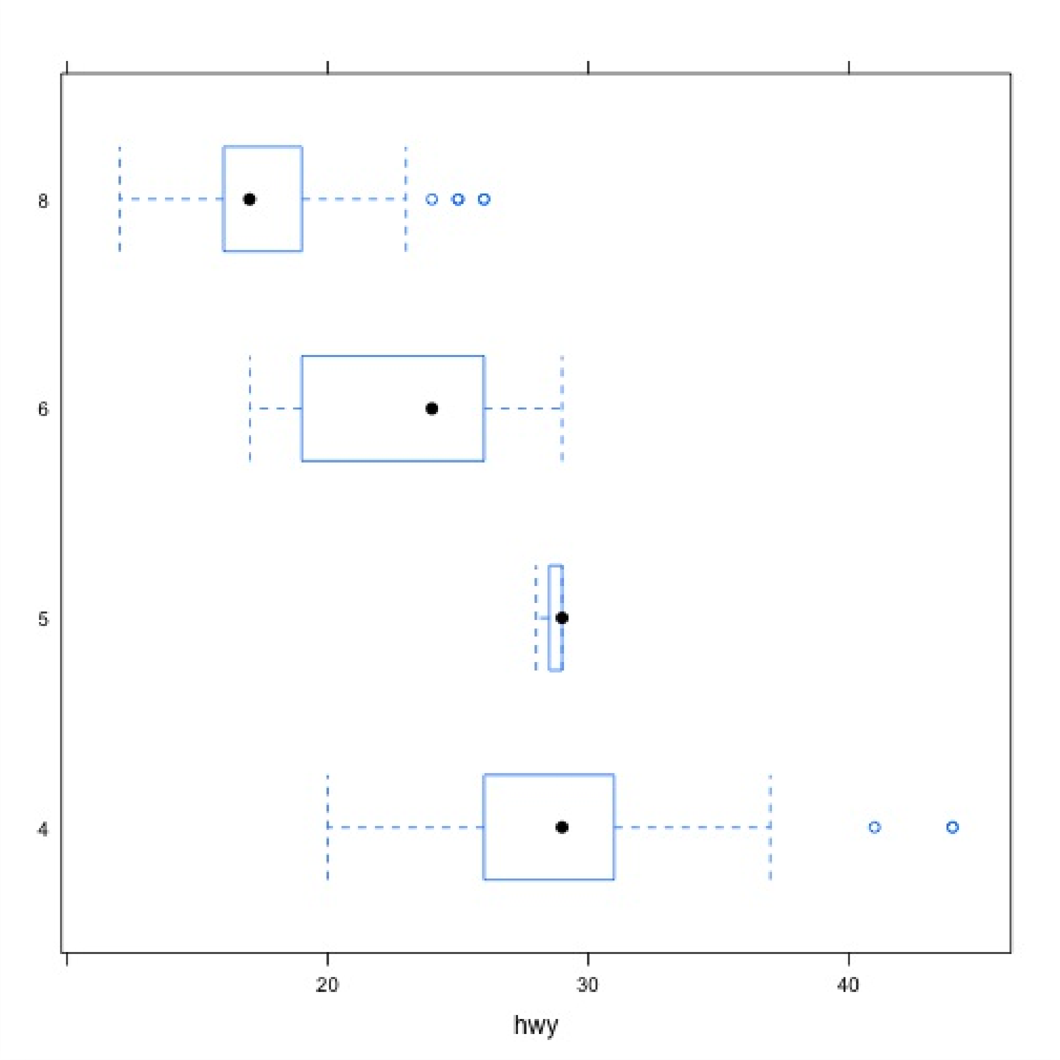
Show comparisons

```
require("ggplot2")
require("lattice")
# head(mpg)

bwplot(~hwy, data = mpg[mpg$cyl == 4, ])
```

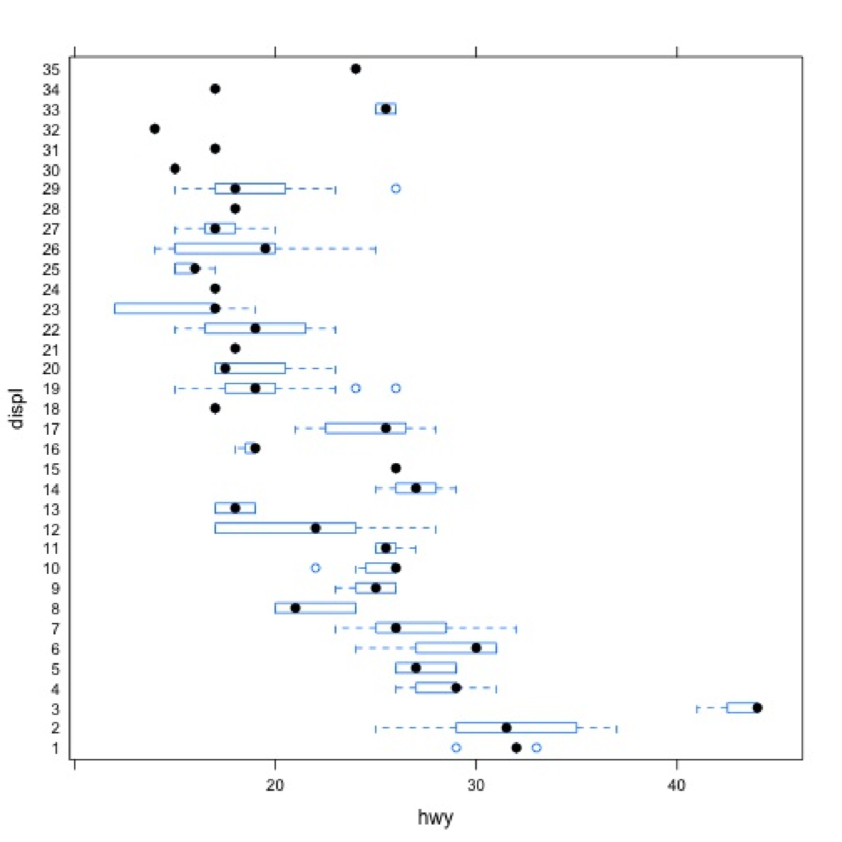


```
bwplot(as.factor(cyl) ~ hwy, data = mpg)
```



Show causality, mechanism, explanation, systematic structure

```r
bwplot(displ ~ hwy, data = mpg)
```

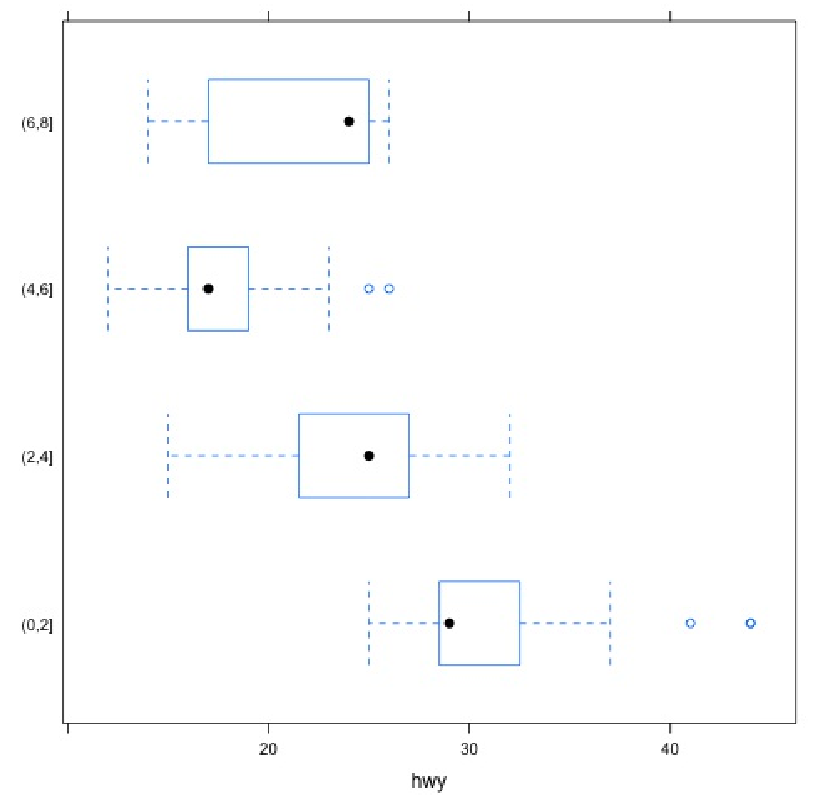


```r
# breaks
displ.int = cut(mpg$displ, breaks = c(0, 2, 4, 6, 8))
```

Doesn't look good so we need to transform the data a little.

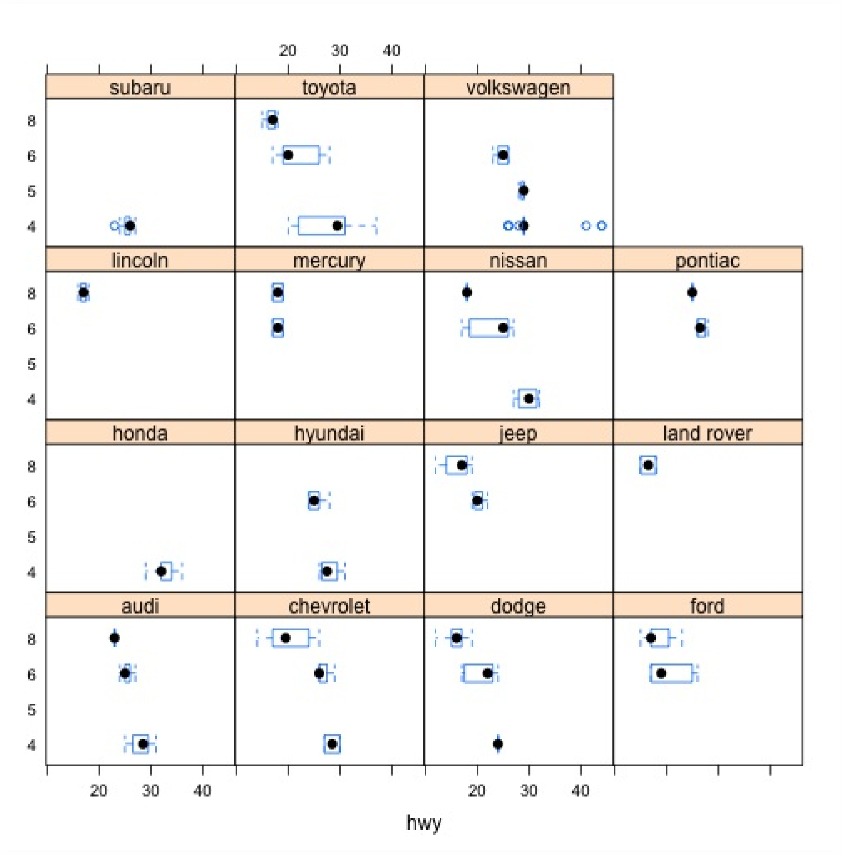
Show causality, mechanism, explanation, systematic structure

```
mpg.new = data.frame(mpg, displ.int)
bwplot(displ.int ~ hwy, data = mpg.new)
```



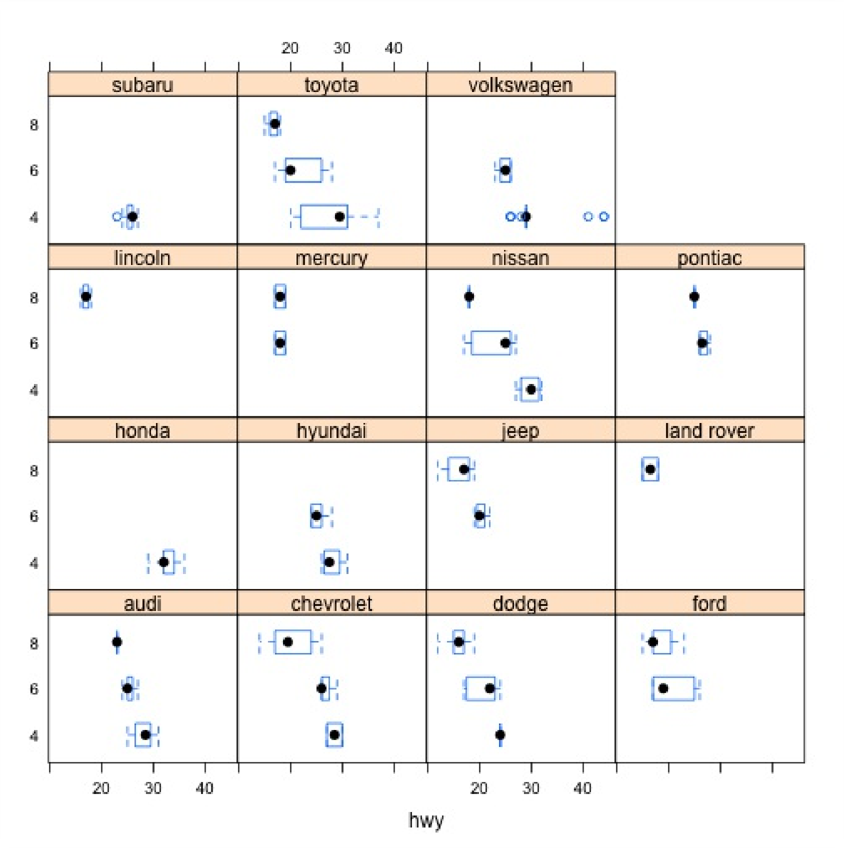
Show multivariate data

```
bwplot(as.factor(cyl) ~ hwy | manufacturer, data = mpg.new)
```



Show multivariate data

```
bwplot(as.factor(cyl) ~ hwy | manufacturer, data = mpg.new[mpg.new$cyl !=
    5, ])
```



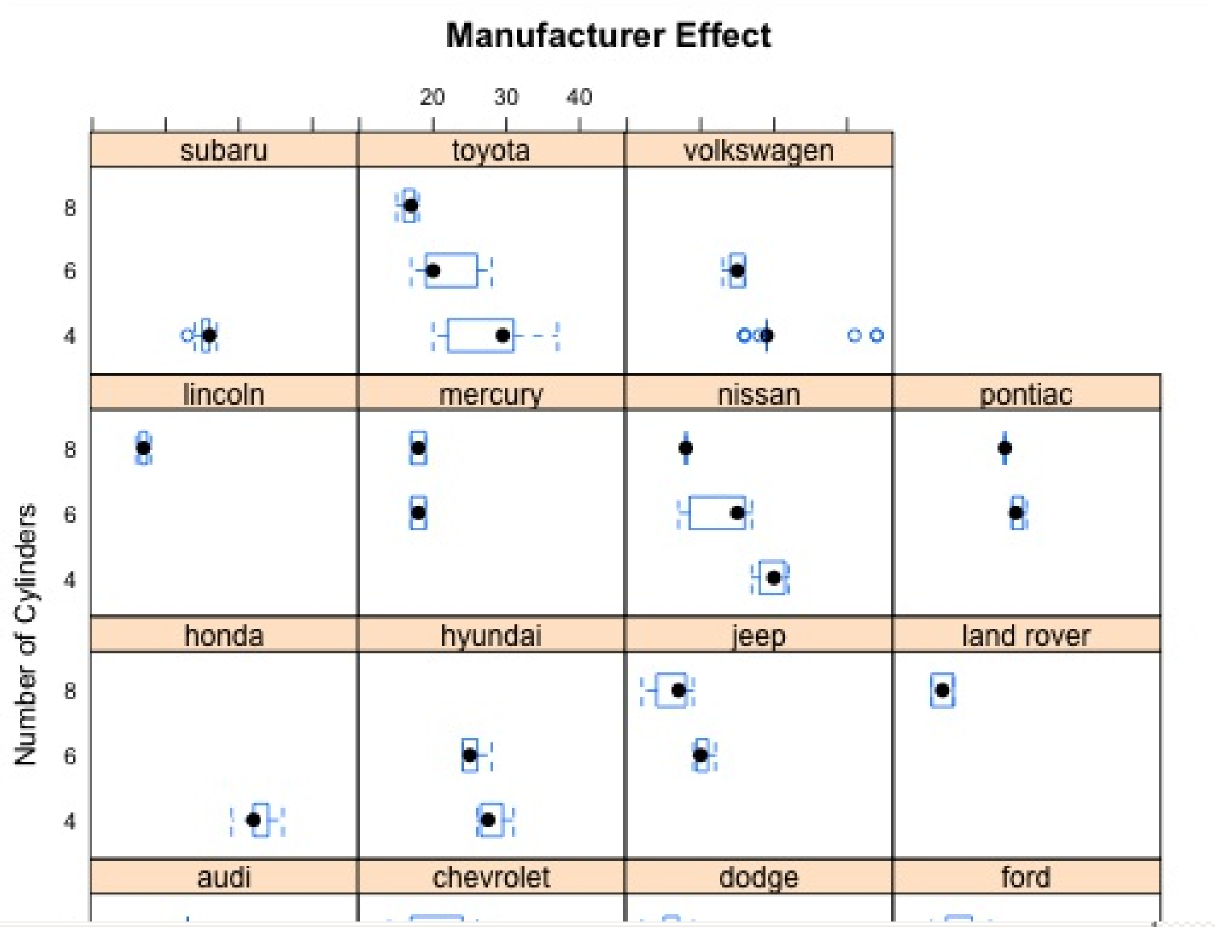
Integrate evidence

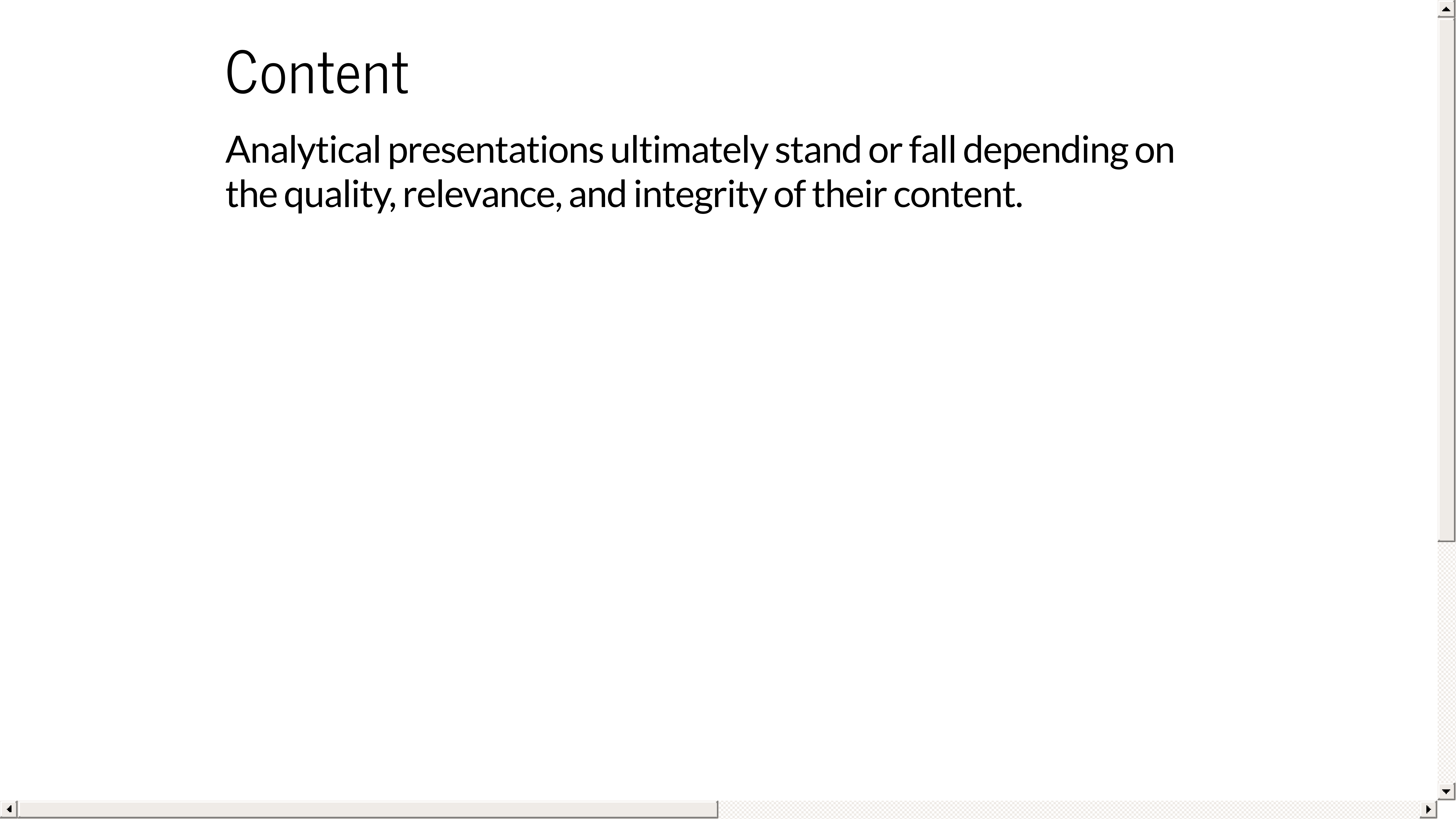
You can also include printed numbers, words, images, and diagrams to tell your story.

Describe and document the evidence

Data graphics should be appropriately documented with labels, scales, and sources. A general rule for me is that a data graphic should tell a complete story all by itself

```
bwplot(as.factor(cyl) ~ hwy | manufacturer, data = mpg.new[mpg.new$cyl !=
    5, ], xlab = "Mpg in Highway", ylab = "Number of Cylinders",
    main = " Manufacturer Effect ")
```





Content

Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.

Summary

- One of the best way to explore data is to visualize it.
- Graphs help us
  - understand and describe the data
  - spot interesting phenomena
  - ask right questions