Exploratory Data Analysis I

Therri Usher

Tuesday, February 24, 2015

Review: Research Question

Throughout the class, I will be working on a data analysis project as well.

I have come up with the following question:

> *Does physical activity differ between black and white older adults?*

Review: Research Question

Consider the following question:

> *Is racial status associated with increased odds of being physically active among non-Hispanic black, non-Hispanic white, and Hispanic older adults aged 65 and older?*

The Main Idea Behind EDA

Now you have cleaned, processed data. So what does it all mean? You can stare at the data, maybe try to read it all... You still probably will not know what the data is saying.

Exploratory data analysis gives us an initial idea of what the data looks like. It can tells us about patterns, relationships, and main points of the data.

What is EDA?

Exploratory data analysis (EDA) is “an approach to analyzing data sets to summarize their main characteristics, often with visual methods”[1].

Almost any method of looking at the data that does not include formal statistical modeling and testing can be perceived as EDA.

Some people even believe some kinds of testing fall under EDA, such as testing the means, proportions, and percentages of two groups.

[1] “Exploratory data analysis.” Wikipedia. http://en.wikipedia.org/wiki/Exploratory_data_analysis

Uses of EDA

Exploratory data analysis allows us to:

- Visualize the data
- Detect mistakes and errors
- Check assumptions of statistical models
- Select the right statistical model
- Explore relationships between variables, including the outcome variable
- Determine which variables should be in the model

Types of EDA

EDA can be broken down into univariate and multivariate (typically bivariate) and graphical and non-graphical:

Univariate non-graphical

- Statistical summaries
- Frequency tables

Univariate graphical

- Histograms
- Boxplots
- Stem-and-leaf plots
- Quantile-normal plots

Types of EDA

Multivariate non-graphical

- Contingency (cross-tabulated) tables
- Correlation and covariance matrices

Multivariate graphical

- Scatterplots
- Boxplots
- Histograms and density plots
- Quantile-quantile plots

Discussion Time

Think about times that you have performed exploratory data analysis. Why did you do so? Did you obtain information about the data? If so, how? If not, why not?

Discuss in your groups and be prepared to share with the class.

What Kind of Information Should I Get From EDA?

The exact information depends on the method of EDA you choose. Overall, your EDA should provide insight into the following questions:

Univariate

- What does the (sample) distribution of the variable look like?
- What is the center and spread of the distribution?
- What is the shape of the distribution
- Does the variable follow a known distribution?

What Kind of Information Should I Get From EDA?

Multivariate

- How do these variables relate to each other?
- Are the variables correlated?
- If correlated, are the variables linearly associated?
- How does the center and spread of their distributions compare?

Overall

- What kind of model should I used to fit the data?
- Does the data fit the assumptions of the model?
- What variables should I include in the model?

Statistical Summaries

- Useful for all data but particularly for continuous
- Provides information about mean, median, variance, quantiles, minimum, and maximum

```
x <- rnorm(100, mean=0, sd=1)
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.1300 -0.7150 -0.0476 -0.1020  0.6590  1.9900
```

```
sd(x)
```

```
## [1] 1.049
```

Statistical Summaries

```
y <- rbinom(100, size=1, prob=0.5)
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.41    1.00    1.00
```

```
var(y)
```

```
## [1] 0.2443
```

```
z <- rpois(100, lambda=10)
summary(z)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3       8      10      10      12      18
```

Frequency and Contingency Tables

- Good for categorical data with few possible values
- Gives an idea of the proportions of the data

```r
table(z)
```

```
## z
##  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
##  5  1  2  7  7  7  9 13 15 19  4  4  2  1  3  1
```

Frequency and Contingency Tables

```r
table(z,y)
```

```
##     y
## z     0  1
##   3   4  1
##   4   0  1
##   5   2  0
##   6   3  4
##   7   3  4
##   8   5  2
##   9   7  2
##   10  8  5
##   11  8  7
##   12 12  7
##   13  2  2
##   14  2  2
##   15  2  0
##   16  0  1
```

Frequency and Contingency Tables

```r
chisq.test(z)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  z
## X-squared = 102.8, df = 99, p-value = 0.3772
```

```r
chisq.test(z,y)
```

```
## Warning: Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  z and y
## X-squared = 15.49, df = 15, p-value = 0.417
```

Frequency and Contingency Tables

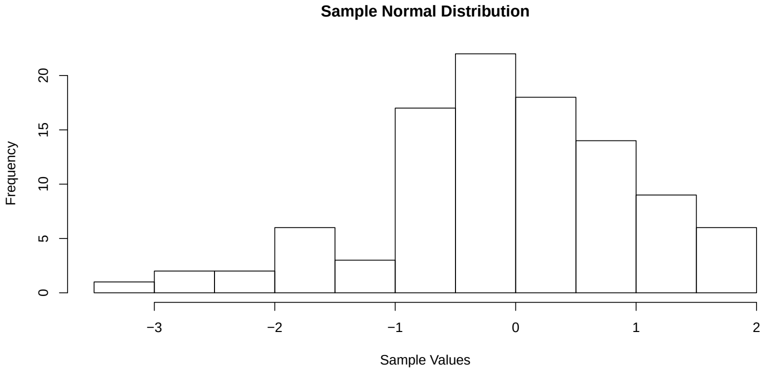
```r
fisher.test(z,y)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  z and y
## p-value = 0.5012
## alternative hypothesis: two.sided
```

Histograms

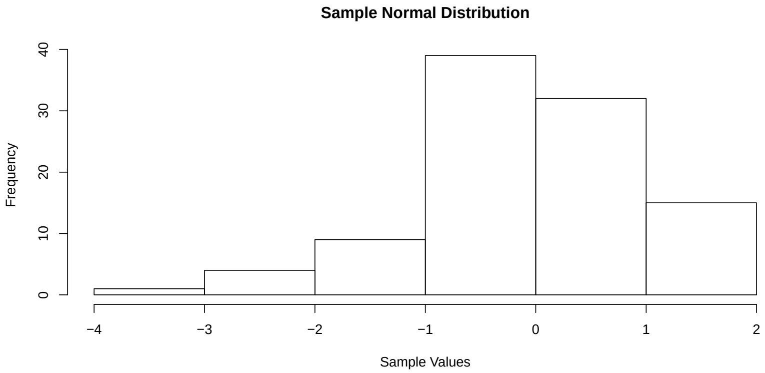
- Good for looking at the shape of the distribution of the variable

```
hist(x, main="Sample Normal Distribution",
     xlab="Sample Values")
```



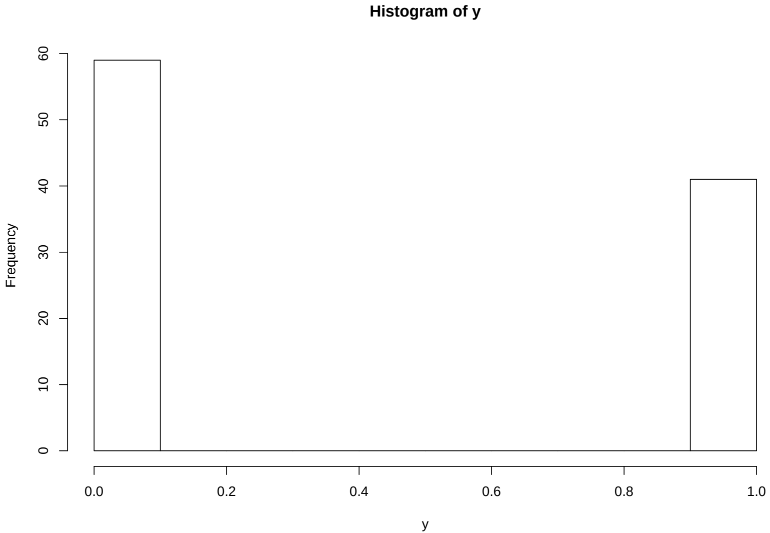
Histograms

```r
hist(x, main="Sample Normal Distribution",
     xlab="Sample Values", breaks=5)
```



Histograms

```
hist(y)
```



Histograms

```
example <- 30:35
example
```

```
## [1] 30 31 32 33 34 35
```

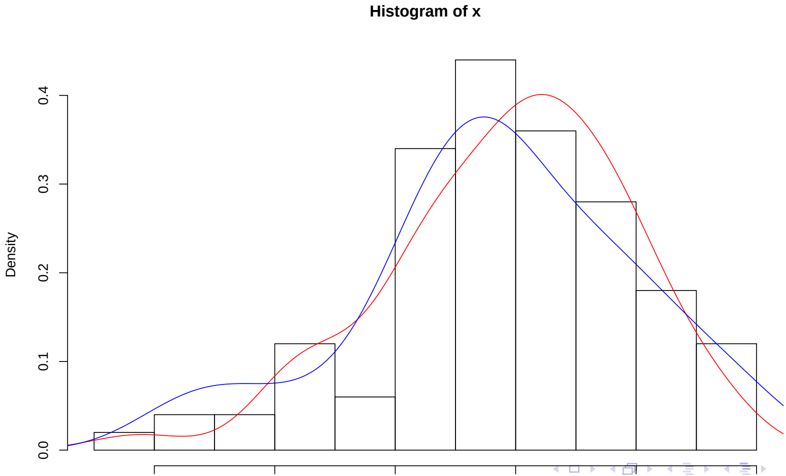
```
which(example==32 | example==34)
```

```
## [1] 3 5
```

```
dat <- as.data.frame(cbind(x, y))
density0 <- density(x[which(y==0)])
density1 <- density(x[which(y==1)])
```

Histograms

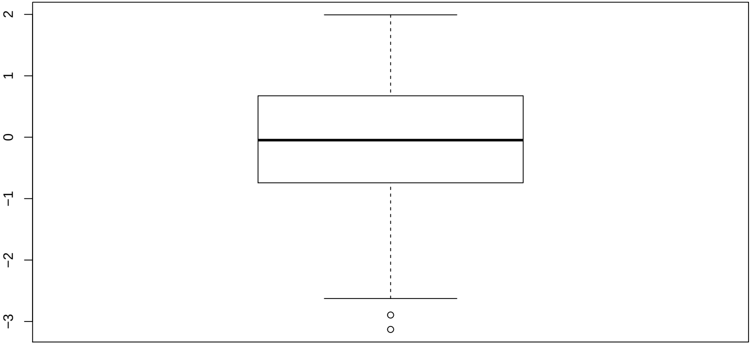
```r
hist(x, freq=FALSE)
lines(density0, col="red")
lines(density1, col="blue")
```



Boxplots

- Provides visuals of the center and spread of a distribution
- Good for continuous data or categorical data with many values

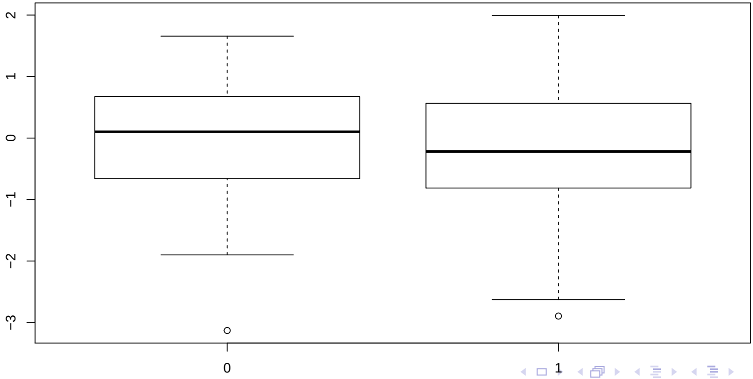
```
boxplot(x)
```



Boxplots

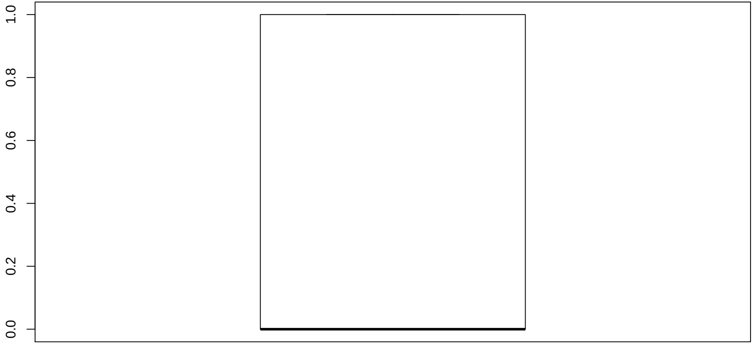
- Can also be used to compare center and spread of multiple distributions

```
y.factor <- as.factor(y)
boxplot(x ~ y.factor)
```



Boxplots

```
boxplot(y)
```



```
# Doesn't look too informative
```

Stem and Leaf Plots

- Can view the shape of the distribution while viewing the actual values
- Good for continuous data or categorical data with multiple values

```r
stem(x)
```

```
##
##   The decimal point is at the |
##
##   -3 | 1
##   -2 | 9630
##   -1 | 998776531
##   -0 | 999999888887766665544443322222221111000
##    0 | 11112223333333455566667778888999
##    1 | 00011123556679
##    2 | 0
```

Stem and Leaf Plots

```r
stem(z)
```

```
##
##   The decimal point is at the |
##
##    2 | 00000
##    4 | 000
##    6 | 00000000000000
##    8 | 0000000000000000
##   10 | 0000000000000000000000000000
##   12 | 00000000000000000000000
##   14 | 000000
##   16 | 0000
##   18 | 0
```

Stem and Leaf Plots

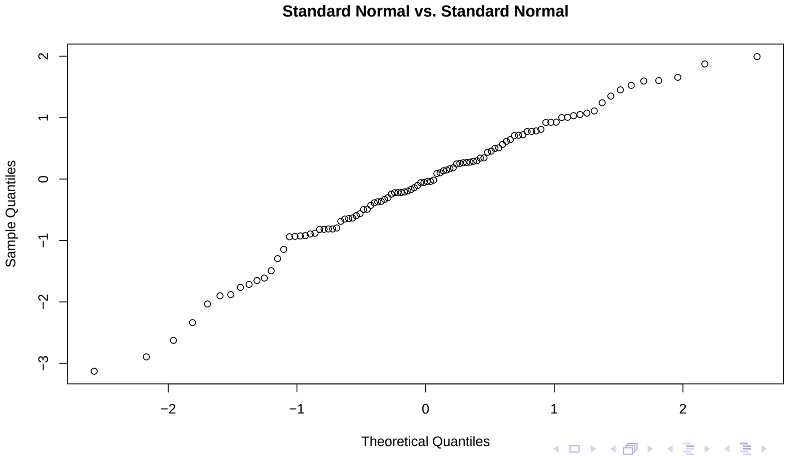
```
stem(y)
```

```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##    0 | 0000000000000000000000000000000000000000000000000
##    2 |
##    4 |
##    6 |
##    8 |
##   10 | 00000000000000000000000000000000000000000
```

Quantile-Normal Plots

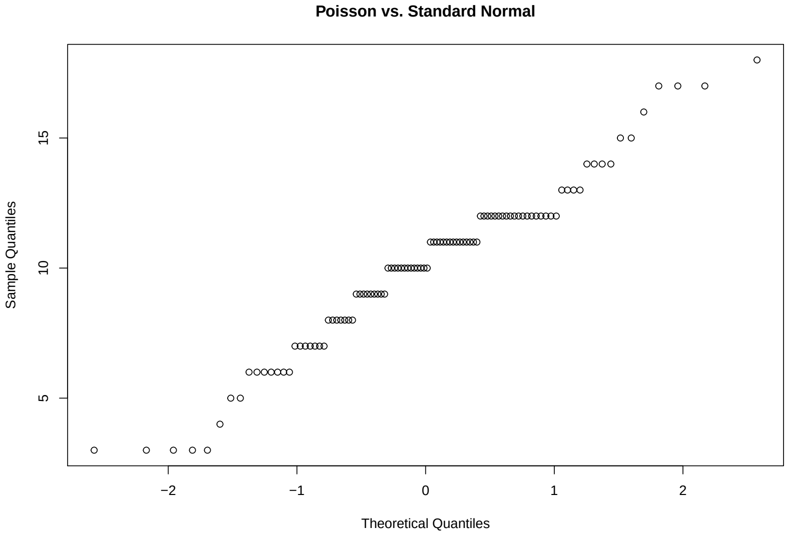
- Compares the distribution of a variable with the standard normal distribution

```
qqnorm(x, main="Standard Normal vs. Standard Normal")
```



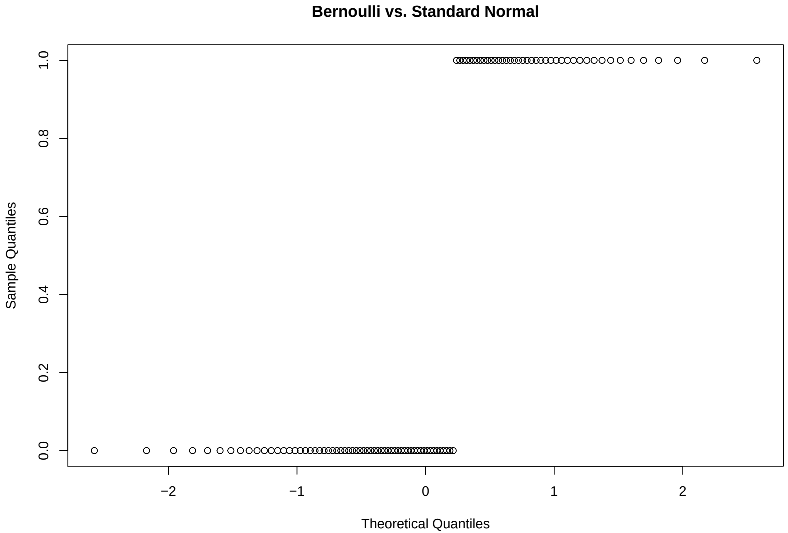
Quantile-Normal Plots

```
qqnorm(z, main="Poisson vs. Standard Normal")
```



Quantile-Normal Plots

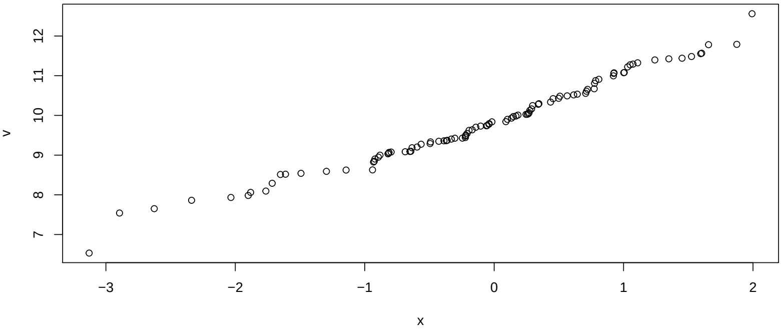
```r
qqnorm(y, main="Bernoulli vs. Standard Normal")
```



Quantile-Quantile Plots

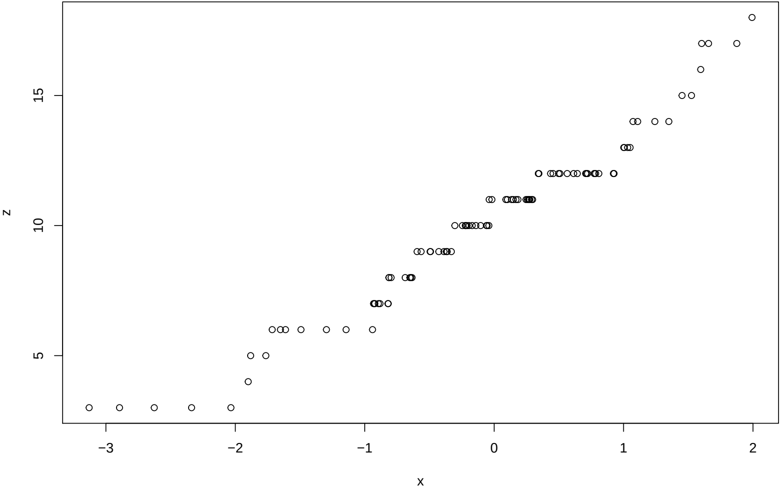
- Used to compare the distributions of two variables

```r
v <- rnorm(100, mean=10, sd=1)
qqplot(x, v)
```



Quantile-Quantile Plots

```
qqplot(x, z)
```



Correlation and Covariance Matrices

- Quantifies correlations and associations between variables
- Good for continuous variables but there are methods for categorical and binary variables

```
x.new <- rnorm(100)
x.linear <- rnorm(100, mean=7, sd=1.75)
x.dat <- as.data.frame(cbind(x, x.new, x.linear))

cov(x.dat, use="complete.obs", method="pearson")
```

```
##                 x   x.new x.linear
## x         1.10014 -0.1500  0.03085
## x.new    -0.14997  0.7805  0.11530
## x.linear  0.03085  0.1153  3.25011
```

Correlation and Covariance Matrices

```
cor(x.dat, method="pearson")
```

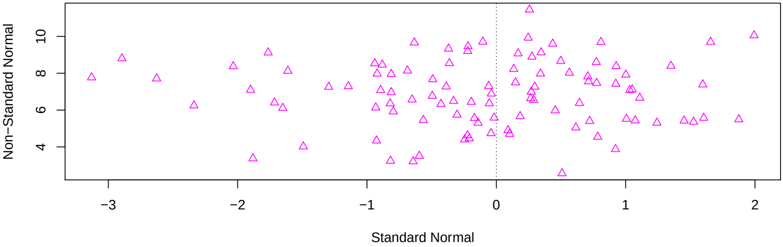
```
##                  x   x.new x.linear
## x          1.00000 -0.1619  0.01632
## x.new     -0.16185  1.0000  0.07240
## x.linear   0.01632  0.0724  1.00000
```

See http://www.statmethods.net/stats/correlations.html for more information.

Scatterplots

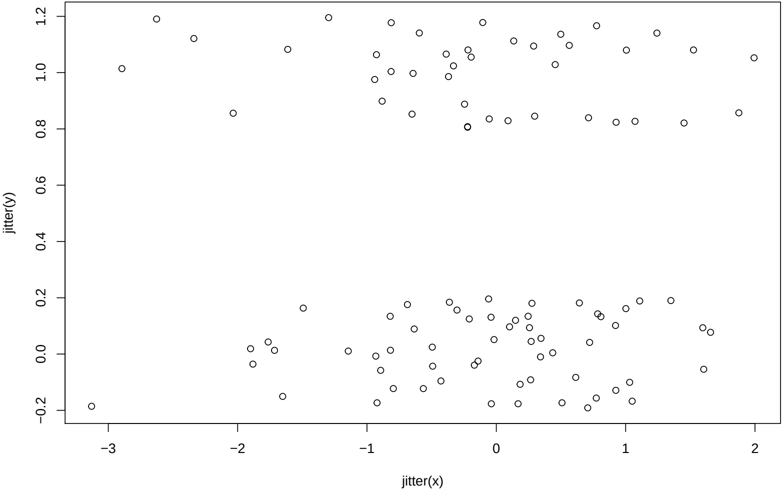
- Allows us to visualize data points for two variables and compare their relationship
- Good for continuous and categorical variables but sometimes used for binary variables

```
plot(x, x.linear, xlab="Standard Normal",
     ylab="Non-Standard Normal", pch=2, col=30)
abline(v=0, lty=3)
```



Scatterplots

```
plot(jitter(x), jitter(y))
```



Options for Scatterplots

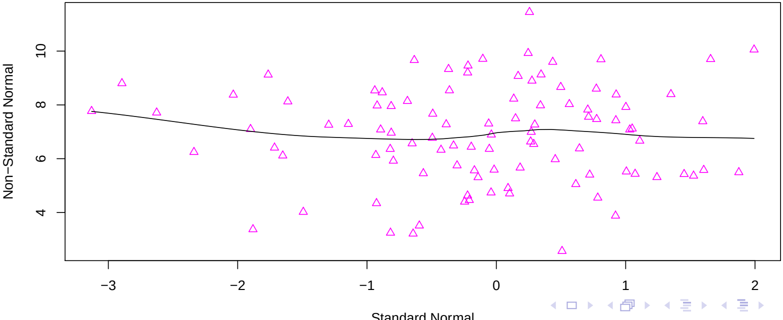
- cex: controls text and symbol size
- pch: controls plotting symbols
- bg: controls fill color of the plotting symbols
- col: controls colors
- lty: controls type (pattern) of lines

See `http://www.statmethods.net/advgraphs/parameters.html` for more information.

Smoothing Curves

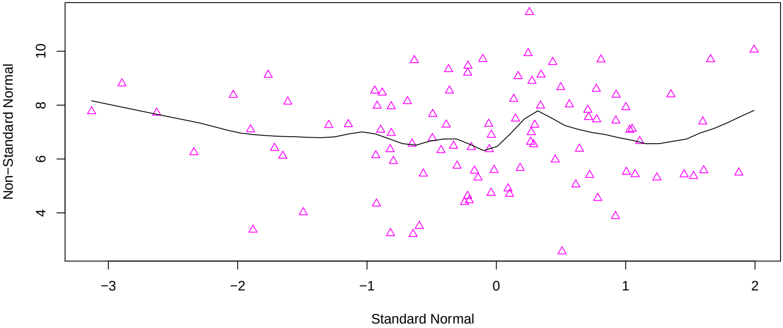
Sometimes a scatterplot can seem like a jumbled mess. Smooth curves are added, which allows the viewer to see the trend in the scatterplot.

```
scatter.smooth(x, x.linear, xlab="Standard Normal",
               ylab="Non-Standard Normal",
               pch=2, col=30, span=2/3)
```

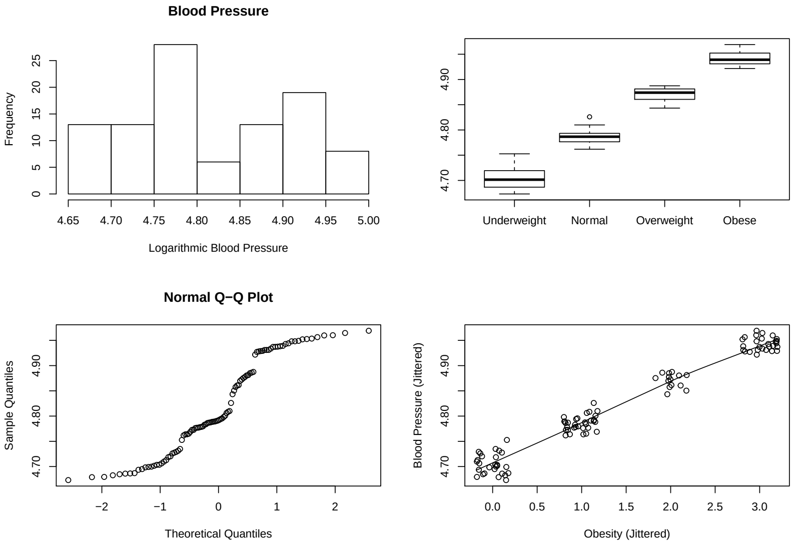


Smoothing Curve

```
scatter.smooth(x, x.linear, xlab="Standard Normal",
              ylab="Non-Standard Normal",
              pch=2, col=30, span=1/4)
```



Active Learning Exercise



Active Learning Exercise

Each group will be assigned to interpret one graph and present their findings to the rest of the class. Each group should be prepared to explain:

- What graph they are interpreting
- What the graph is used for in order to understand the data
- What the graph says about this particular data

Once all groups have presented, the class will work together to answer the following questions:

- Is obesity status related to blood pressure?
- Is a linear regression model suitable for this data?

Multivariate EDA

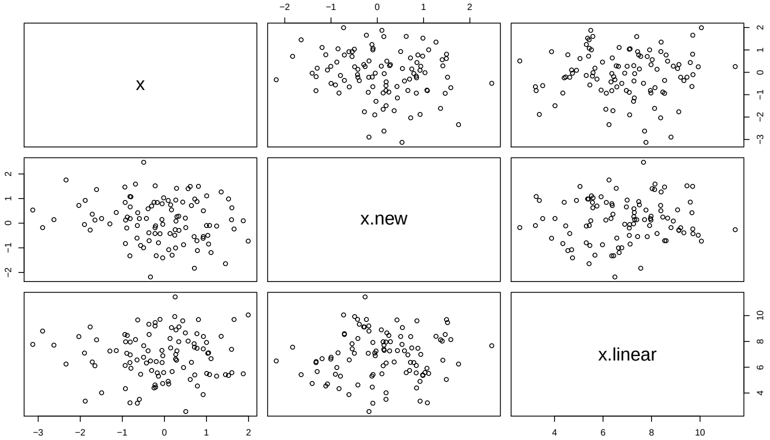
Notice that most of the EDA presented only looks at two variables at a time. There are methods to look at more than two variables at a time, such as:

- Grid scatterplots
- ANOVA
- Spaghetti plots

Grid Scatterplots

- Create pairwise scatterplots of all variables in a matrix or data frame

```
plot(x.dat)
```



ANOVA

Analysis of variance (ANOVA) allows us to compare group means by looking at the variation in the data, between-group and within-group variation.

Ho: all group means are the same

Ha: at least 2 groups have different means

ANOVA

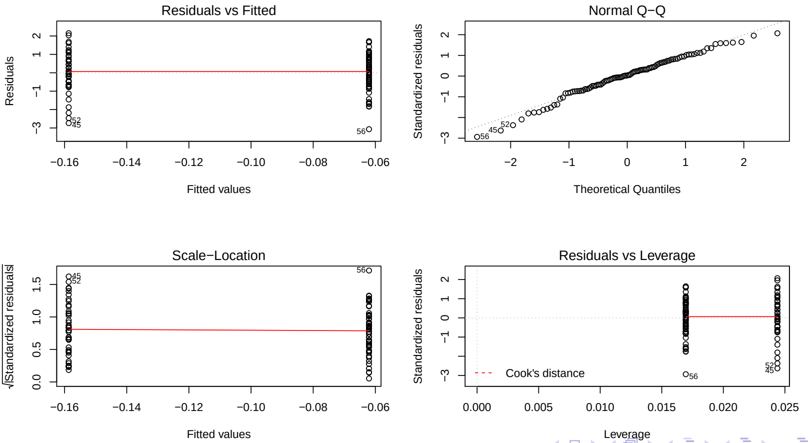
```r
fit <- aov(x ~ y.factor)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## y.factor     1    0.2   0.226     0.2   0.65
## Residuals   98  108.7   1.109
```

ANOVA

Assumptions: Same as linear regression.

```r
par(mfrow=c(2,2))
plot(fit)
```



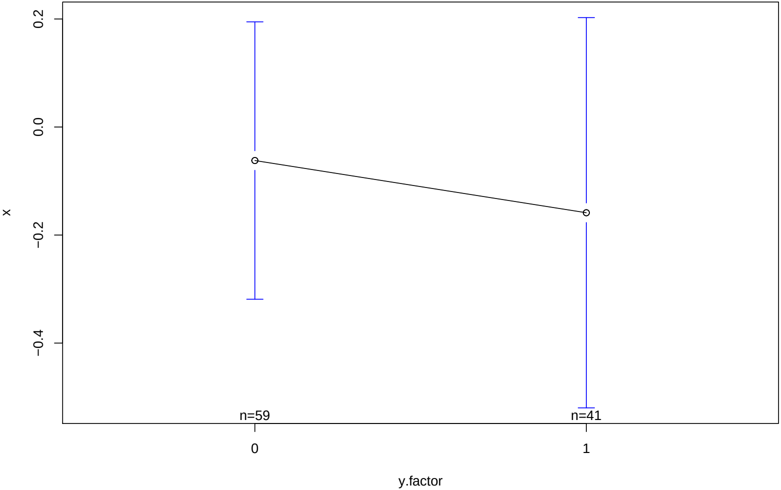
ANOVA

```r
library(gplots)
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
```

ANOVA

```
plotmeans(x ~ y.factor, p=0.95)
```



Multivariate EDA

Other methods are more advanced:

- 3D plots
- Principal component analysis
- Factor analysis
- Cluster analysis

ggplot2 package

What is ggplot2?

ggplot2 is an R package created by Hadley Wickham designed for making statistical graphics.

It differs from other graphics packages in that its grammar is based on the Grammar of Graphics.

Grammar of Graphics

The grammar of graphics argues that a statistical graphic is a mapping from data to aesthetic attributes (ex: color, shape, and size) to geometric objects (ex: points, lines, bars).

The plots are created on a specific coordinate system.

Faceting can be used to generate the same plots for subsets of the data.

Grammar of Graphics

As a result, a graphic is made of the combinations of these independent components:

- data
- aesthetic mappings
- geometric objects
- statistical transformations
- scales map
- coordinate system
- faceting specification

Advantages of ggplot2

- Users are not limited to a set of pre-specified graphs.
- Users do not have to worry about details, such as drawing legends.
- Graphics are created in layers.
    - You can make graphs you have not even thought of yet.
- Hadley Wickham

Disadvantages of ggplot2

- It requires a new way of thinking.
- The grammar does not specify all parts of the graph (ex: background color and font size)
- It cannot be used for interactive graphics.
- It's still a computer program.

Basics of ggplot2

The basic way to make a plot is with the qplot() command. Just specify the variables you want to plot.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

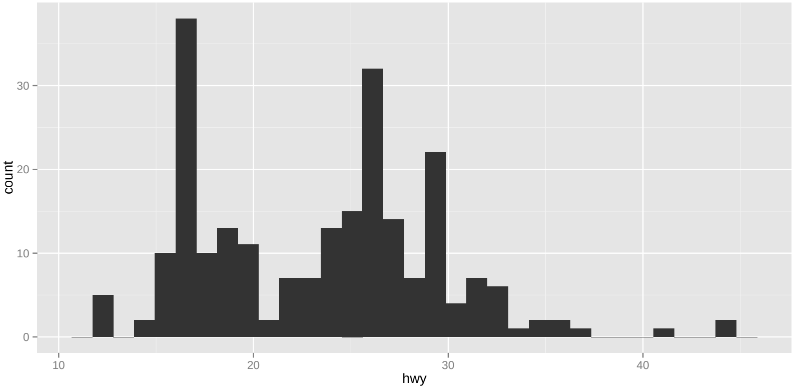
```
data(mpg)
head(mpg)
```

```
##   manufacturer model displ year cyl      trans drv cty
## 1         audi    a4   1.8 1999   4   auto(l5)   f  18
## 2         audi    a4   1.8 1999   4 manual(m5)   f  21
## 3         audi    a4   2.0 2008   4 manual(m6)   f  20
## 4         audi    a4   2.0 2008   4   auto(av)   f  21
## 5         audi    a4   2.8 1999   6   auto(l5)   f  16
## 6         audi    a4   2.8 1999   6 manual(m5)   f  18
```

Basics of ggplot2

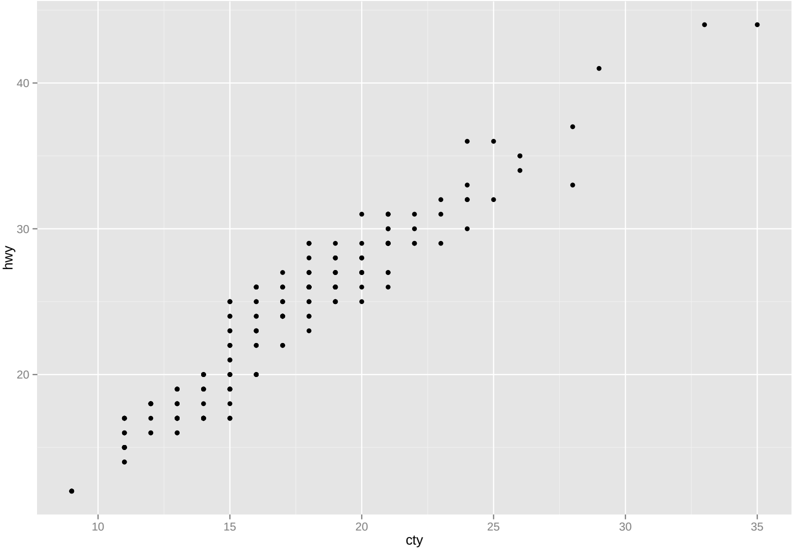
```
qplot(hwy, data=mpg)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth
```



Basics of ggplot2

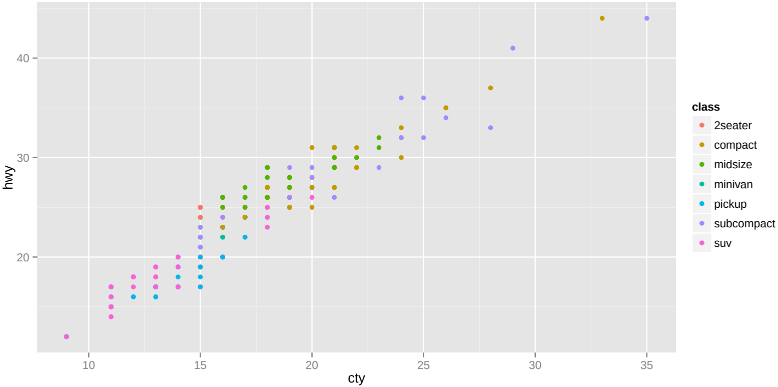
```
qplot(cty, hwy, data=mpg)
```



Aesthetic Attributes

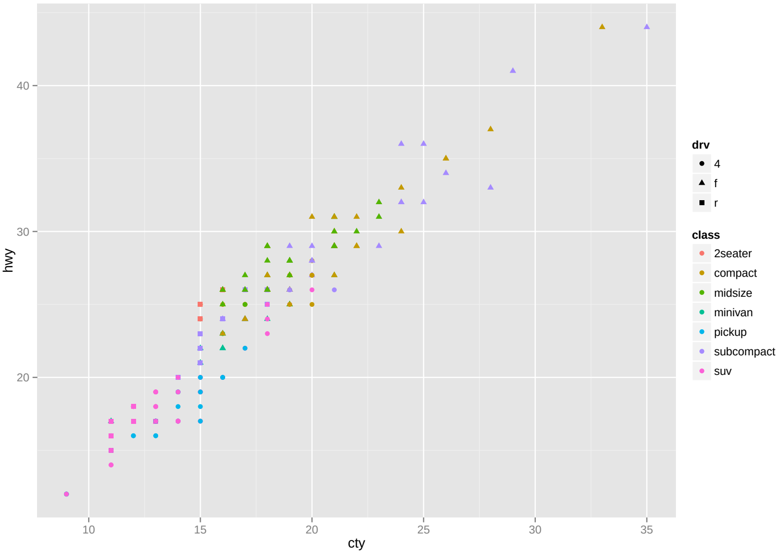
ggplot2 maps a data variable to an attribute of the graph

```
qplot(cty, hwy, data=mpg, colour=class)
```



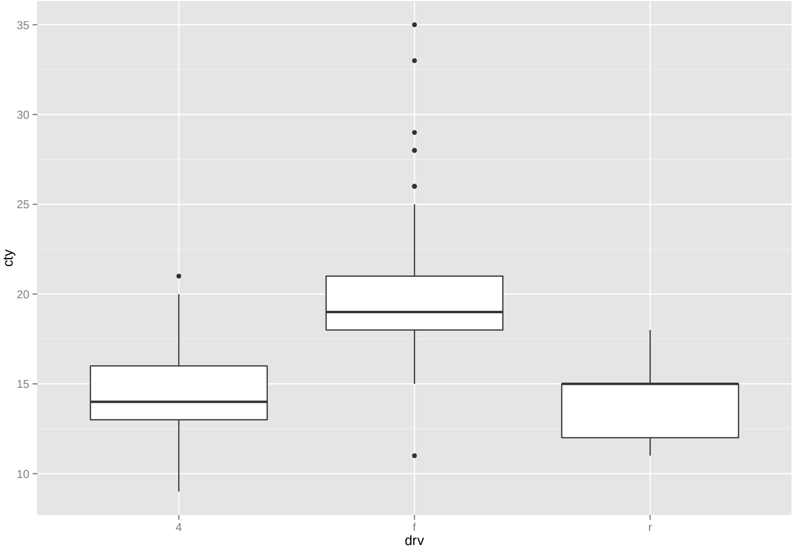
Aesthetic Attributes

```
qplot(cty, hwy, data=mpg, colour=class, shape=drv)
```



Geometric Objects

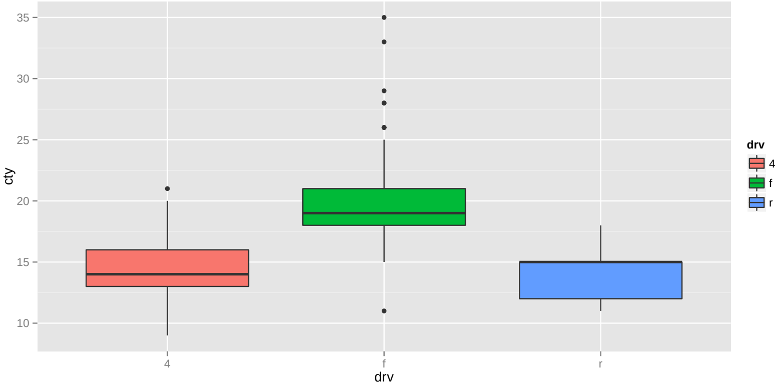
```
qplot(drv, cty, data=mpg, geom="boxplot")
```



Geometric Objects

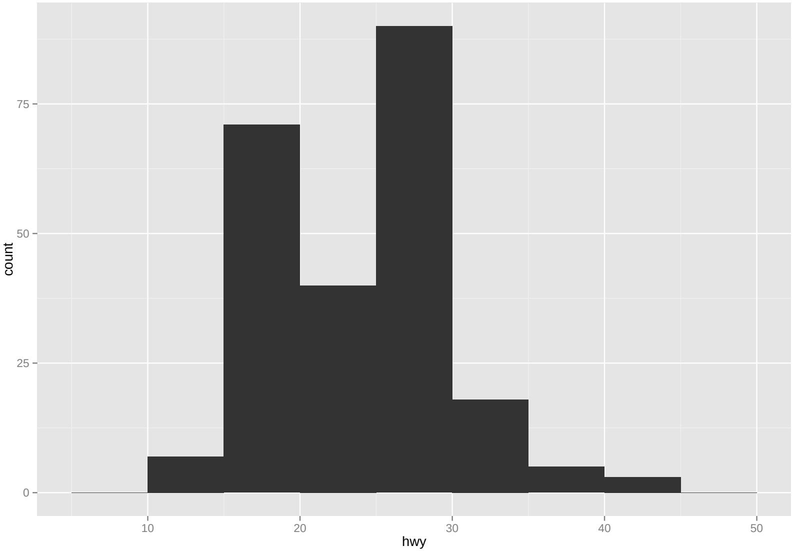
geom can take the values “point”, “smooth”, “boxplot”, “line”, and “path”, “histogram”, “freqpoly”, “density”, and “bar”

```
qplot(drv, cty, data=mpg, geom="boxplot", fill=drv)
```



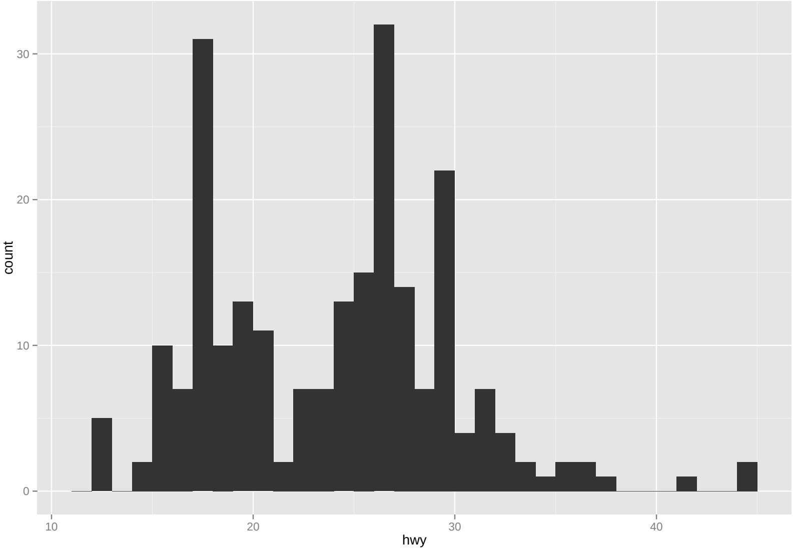
Geometric Objects

```
qplot(hwy, data=mpg, geom="histogram", binwidth=5)
```



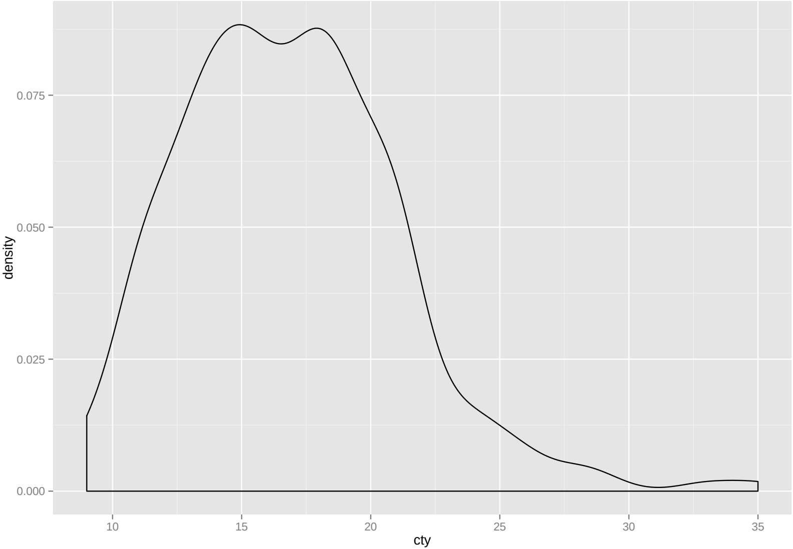
Geometric Objects

```
qplot(hwy, data=mpg, geom="histogram", binwidth=1)
```



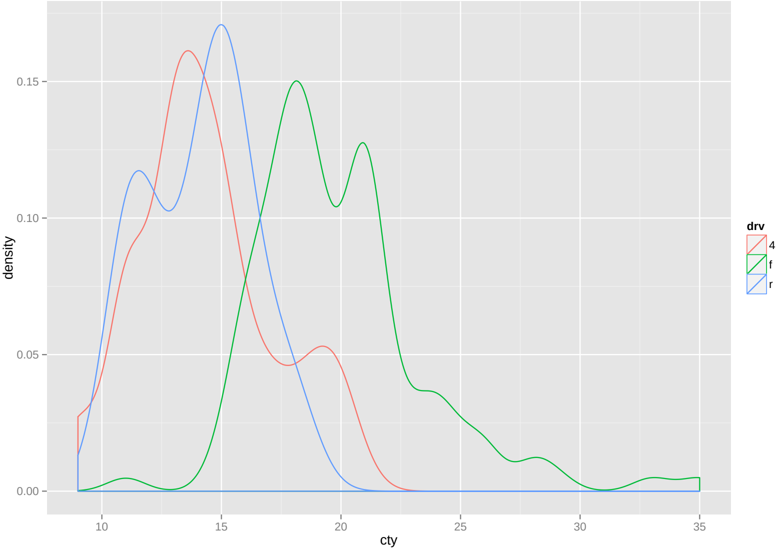
Geometric Objects

```
qplot(cty, data=mpg, geom="density")
```



Geometric Objects

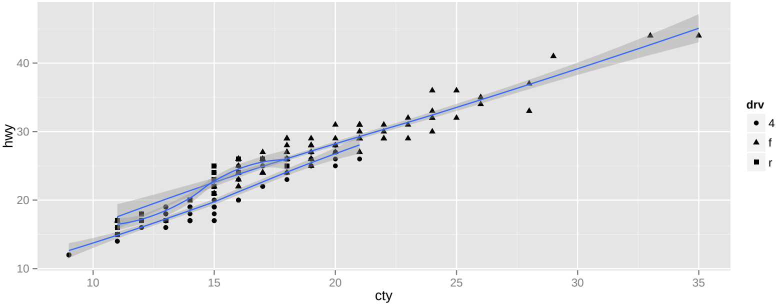
```
qplot(cty, data=mpg, geom="density", colour=drv)
```



Adding a Smoother

```
qplot(cty, hwy, data=mpg, geom=c("point", "smooth"),
      shape=drv, span=1)
```

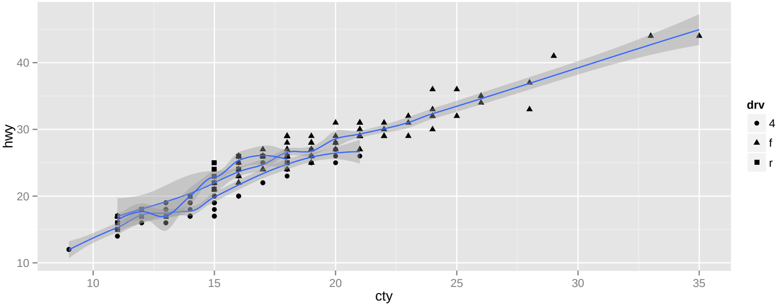
```
## geom_smooth: method="auto" and size of largest group is
```



Adding a Smoother

```
qplot(cty, hwy, data=mpg, geom=c("point", "smooth"),
      shape=drv, span=0.5)
```

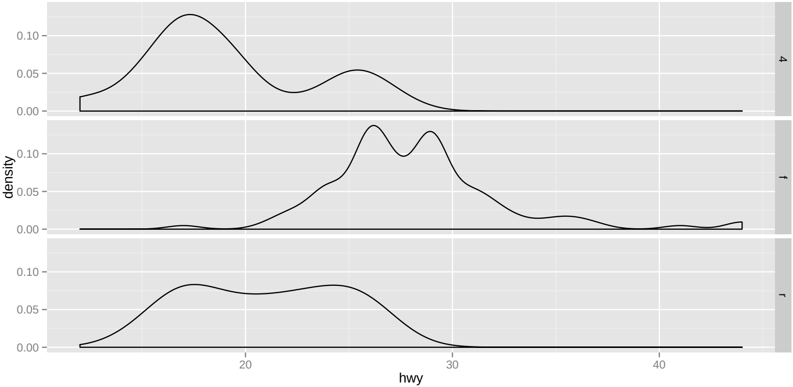
```
## geom_smooth: method="auto" and size of largest group is
```



Faceting

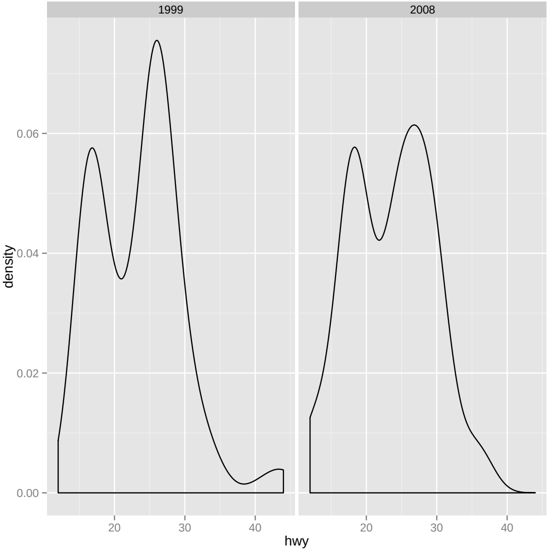
Faceting allows us to plot subsets of the data on separate graphs placed on a grid. We do this by specifying facets = row_var ∼ col_var

```
qplot(hwy, data=mpg, geom="density", facets=drv~.)
```



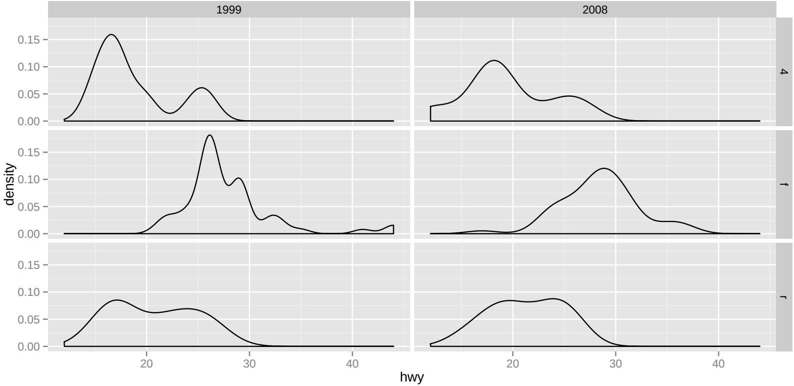
Faceting

```
qplot(hwy, data=mpg, geom="density", facets=.~year)
```



Faceting

```
qplot(hwy, data=mpg, geom="density", facets=drv~year)
```



Discussion Time

How does ggplot2 compare to the basic graphics functions in R?

Discuss some examples in your group.

Tip of the Iceberg

These slides do not begin to cover all the capabilities of ggplot2. If you are really interested, I recommend downloading Hadley Wickham's ggplot2 book through Springer Link.

Things to Keep in Mind

- You can still use the arguments used for the basic plot function, such as main, xlab, and ylab.
- The function qplot is not generic, meaning that you cannot pass an object through it and get a default plot.
- Layers are an integral part of ggplot2. I recommend reading Chapter 3 of Hadley Wickham's book.

Active Learning Activity

Use the data set called “diamonds” that comes in the ggplot2 package.

```r
require(ggplot2)
data(diamonds)
```

We are interested in modeling the price of a diamond using certain characteristics of the diamond. Perform exploratory data analysis using the variables in the data set.

Active Learning Activity

The following are some (not all) of the questions you should try to answer:

- Describe the distribution of the prices and the other variables.
- What variables seem to relate to the price of a diamond?
- Do any of the non-price variables relate to each other?
- What variables should be included in a modeling of price?