Intro to Data Analytics and Visualizations

Lecture 8 – Exploring Data
Fall 2014, September10

Outline

1. Why explore before modeling?
2. Tools used for exploration
3. Summary statistics
4. Graphics and Visualizations

Why explore?

- To spot problems with the data and derive a first feel for what is going on!
- Potential problems with data:
  - Missing variables
  - Missing observations
  - Missing entire subsets
  - Bad values(dirty/inconsistent)
  - Variables need transformation

Tools Used for Exploration

- Summary statistics (descriptive statistics)
- Visualizations (graphics)

Visualizations to Explore Data

- Complementary to numerical summaries;
- Can sometimes spot more issues with the data;
- Give an early feel for some relationships between variables;

Tips for Good Visualizations

- Build it then remove anything non-essential;
- Use colors;
- Make it easy to interpret;
- Avoid background patterns and colors;
- Avoid unnecessary or disorganized text;
- Should convey a lot of information; the message should be clear;
- Pick the right type of graphical display (e.g. pie charts best avoided).

How to Build Visualizations with R

- The ggplot2 package! (there are many others)

- With any new package, that contains a few R functions, before you can use the functions, you need to:

- install the package with:

```
>install.packages(“ggplot2”)
```

-load package functions with:

```
> library(“ggplot2”)
```

-see all the functions available in the package with:

```
>help(package = ggplot2)
```

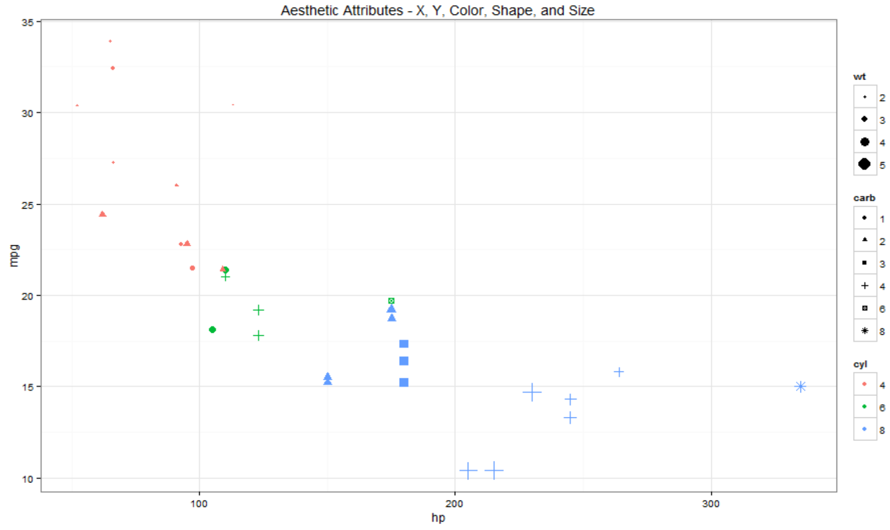
ggplot2 (“grammar of graphics plot 2”

- ggplot2 is an R package for producing statistical graphics

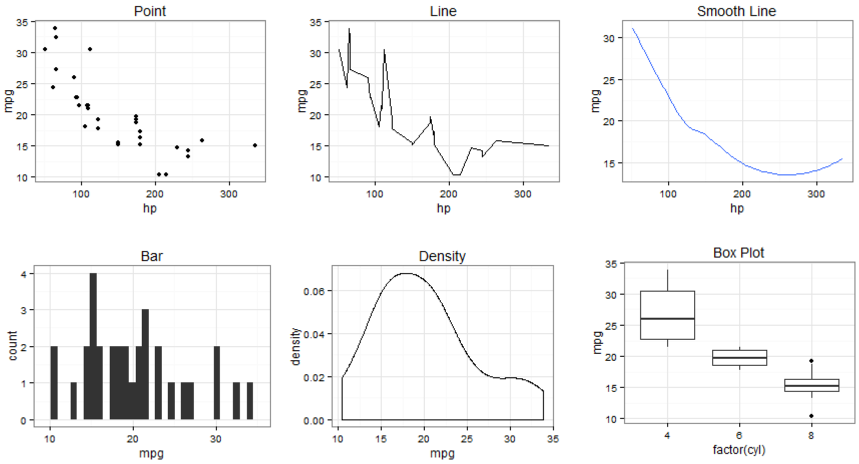
- A statistical graphic is a mapping from data to **aesthetic attributes** of **geometric objects**. The plot may contain **statistical transformations** of the data and is drawn on a specific **coordinate system**. **Faceting** can be used to generate the same plot for different subsets of the data.

aesthetic attributes	position, color, shape, size
geometric objects	points, lines, bars
statistical transformations	binning, counting, regression
coordinate system	Cartesian, polar, latitude/longitude (maps)
faceting	latticing

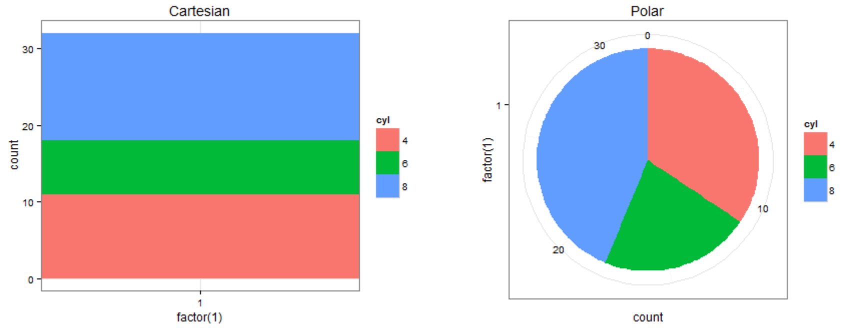
What are aesthetic attributes?



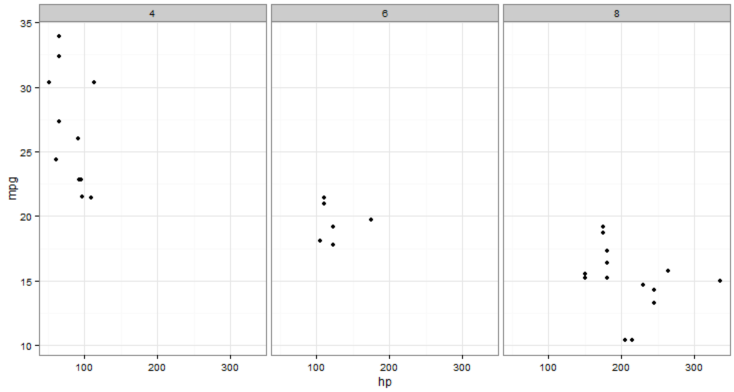
What are geometric objects and statistical transformations?



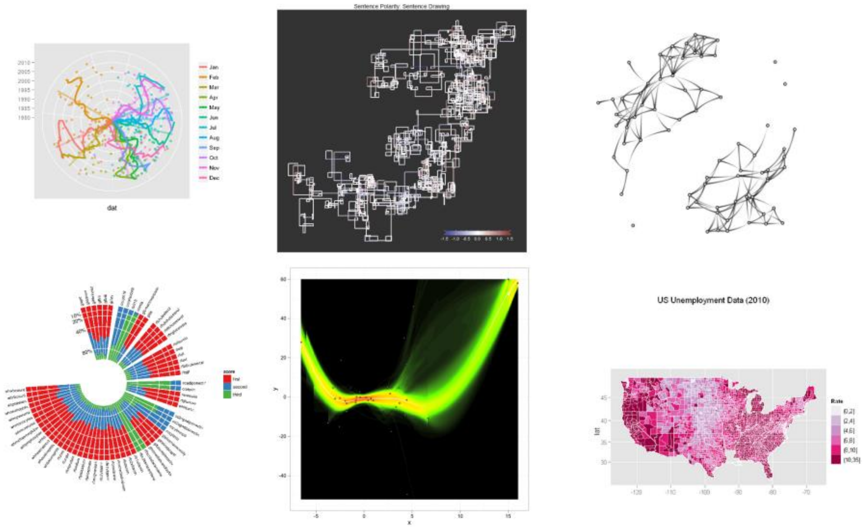
What exactly does coordinate system mean?



What is faceting?



Interesting plots created with GGPLOT2



Common Types of Visualizations

Visualizations can be created to explore one variable or multiple (typically two) variables.

One Variable

- Histograms
- Density plots
- Bar charts

Two Variables

- Line plots
- Scatter plots
- Bar charts

One Variable

Histograms

- Group observations into (bins) using a numeric variable

E.G: By height (51-55in., 56-60, 61-65 etc)

- Build a bar for each category to show how many people fall in each category

E.G: 10 people in 51-55; 15 people in 56-60 etc)

Look at this distribution, what is a typical height?

Density Plots

- Displays similar info with a histogram.
- Can think of it as a smooth curve fit to the histogram.
- It alleviates the histogram's burden of picking the right number of bins

Bar Charts

- For categorical (factor) data
- Bar height shows the number of observations in each category (level) of the factor
- Can be vertical or horizontal