Visualizing Data: Customization with ggplot2

Data Science 101

Stanford University, Department of Statistics
ggplot2: Customizing graphics in R

ggplot2 by RStudio’s Hadley Wickham and Winston Chang offers a highly versatile ‘grammar of graphics’.

- Basic statistical plots are typically no more difficult with ggplot2 than ‘base R’ graphics, especially if qplot (‘quick plot’) is used.
- Some graphics, like ‘faceted’ graphics that make similar plots for each factor level, are considerably easier with ggplot2, while others (like the ‘violin’ plot) simply don’t exist in base R.
- Publication-quality graphics are usually easier with ggplot2.
- Today, most (if not all) active development of graphics in R builds from ggplot2. Examples of recent libraries:
  - ggmap (integrates with Google Maps)
  - ggridges (support for many density plots)
  - gganimation (data animation)
gg basics

- ggplot requires a data.frame.
- aes (as in ‘aesthetics’) is used to specify x and/or y
- Additional features are added (literally using +) to layer on additional function calls, including a variety of statistics

```r
ggplot(mtcars, aes(x=factor(cyl))) + geom_bar()
```
gg coding style

ggplots are often saved as objects to which additional features are added one or two at a time.

```r
# ggplot with custom style

# Define the plot
g <- ggplot(mtcars, aes(x=factor(cyl))) + geom_bar()

# Add a theme with larger text

# Customize axis labels and title

# Apply minimal theme

g
```

```
Cars

Cylinders

N

Cars

0
5
10
4 6 8
Cylinders

N

Cars

0
5
10
4 6 8
Cylinders

N

Cars

4
6
8
```
this grammar of graphics enables endless variety

source: Google Image 4/8/2017
Outline

The slides that follow contain `ggplot2` examples of:

- Histograms
- Box Plots
- Violin Plots
- Scatter Plots
- Loess (Smoothed Trend Lines)
- Density Plots
- Faceting

To avoid clutter, recurring coding details (like font size) are omitted after their first appearance.
Histograms

```r
ggplot(airquality, aes(x=Ozone)) + geom_histogram(binwidth=10)
```

Daily Ozone Levels
New York, May–Sept 1973
Histogram, density example

- Setting the y aesthetic to ..density.. makes the sum of (width of bar * height of bar) add to 1.
- The unit becomes % per unit/100 (% per ppb/100 below).

```r
ggplot(airquality, aes(x=Ozone)) + geom_histogram(binwidth=10, aes(y=..density..))
```

Daily Ozone Levels
New York, May–Sept 1973
Information displayed in histogram

- Density: The height of the bar tells how many observations there are for one unit on the horizontal scale.

- For example, the highest **density** is around 15-25 ppb as $2.5 \times 10 = 25\%$ of all (non-missing) observations have values in the range 15-25 ppb.

- Thus it also displays **relative frequency**

Daily Ozone Levels
New York, May–Sept 1973
ggplot(mtcars, aes(x=factor(cyl), y=mpg)) + geom_boxplot()
Violin plot of Bay Area weather

```r
ggplot(SFO2013,
      aes(x = factor(Month), y = Max_TemperatureF)) +
geom_violin()
```

Daily High at SFO in 2013

Source: library(weatherData)
Scatter plot

```r
ggplot(GaltonFamilies,
      aes(x=midparentHeight, y=childHeight)) + geom_point()
```

Galton's Famous Height Data
N = 934 children in 205 families, Galton (1886)
Adding a Smoothed Trend Line

Sometimes we want to add a trend line (details on how this works will come later in the course).

```r
ggplot(GaltonFamilies, aes(x=midparentHeight, y=childHeight)) + geom_point() + geom_smooth()
```

Galton's Famous Height Data
N = 934 children in 205 families, Galton (1886)
Joint Density Plots

```r
ggplot(GaltonFamilies, aes(x=midparentHeight, y=childHeight)) +
geom_density2d()
```

Galton's Famous Height Data
N = 934 children in 205 families, Galton (1886)
Scatter plot

If there are many points then one needs to modify the **opacity** with the alpha parameter to be able to see anything.

```r
ggplot(flights, 
aes(x=as.numeric(dep_time_sched), y=dep_delay)) + 
geom_point(alpha=.3, size=.5)
```

**Delay at NYC Area Airpots**

source: library(nycflights13)
Scatter plots with more than two variables

For multiple variables, add size, shape, and color to scatters.
Example: Scatter Plot with Four Variables

A single graphic displaying fuel economy (mpg) by cylinders (cyl), weight (wt), and acceleration (qsec).

```
library(ggplot2)

ggplot(mtcars, aes(x=wt, y=mpg, size=qsec, col=factor(cyl))) + geom_point() + labs(x="Weight", size="Acceleration", col="# Cylinders")
```
Conditional Density Plots

\texttt{ggplot(GaltonFamilies, aes(x=childHeight, col=gender)) + geom_density()}

Galton's Famous Height Data
N = 934 children in 205 families, Galton (1886)
Small multiples

- **Small multiples** refers to a series of charts using similar scales and axes, arranged in a lattice or grid.

At the heart of quantitative reasoning is a single question: **Compared to what?** Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives. For a wide range of problems in data presentation, **small multiples are the best design solution** (Tufte, *Envisioning Information*, p67).

- The pair plot is, loosely speaking, a **small multiple** design
- More generally, we can produce small multiples in ggplot by **faceting**
Suppose Galton suspects that genetics aren’t the only determinant of height and that childhood nutrition is very important too. Other things being equal, he imagines that it is harder to feed large families well, so he decides to facet a histogram of height by the number of children in the family...
Faceting Child Height by Family Size

ggplot(GaltonFamilies) +
  geom_bar(aes(childHeight, fill=gender)) +
  facet_grid(children ~.)
ggplot(GaltonFamilies, 
    aes(x = factor(children), y = childHeight)) + 
geom_violin()
Summary

ggplot2 is a highly customizable ‘grammar of graphics’ that enables one to make publication-quality graphics in keeping with the recommendations of visualization experts like Tufte.

- Histograms
- Box Plots
- Violin Plots
- Scatter Plots
- Loess (Smoothed Trend Lines)
- Density Plots
- Faceting