Original Slides

“Intro to R and ggplot2”
by Hadley Wickam, creator of ggplot2
Rstudio
```r
plot(displ, hwy, data = mpg)
```

```
> library(ggplot2)
Loading required package: reshape
Loading required package: plyr

Attaching package: 'reshape'

The following object(s) are masked from 'package:plyr':

  rename, round_any

Loading required package: grid
Loading required package: proto
> plot(displ, cyl, data = mpg)
> plot(displ, hwy, data = mpg)
> ```
Rstudio

Console - run code here

```r
# Load ggplot2 package
library(ggplot2)

# Load reshape package
library(reshape)

# Load pply package
library(plyr)

# Attach reshape package
library(reshape)

# Plot displ vs hwy
qplot(displ, hwy, data = mpg)
```

![R Studio IDE](image)
Rstudio
Output - plots and help

```r
1 aplot(displ, hwy, data = mpg)
```

Dashboard with a plot showing a relationship between `displ` and `hwy`.
Rstudio
Editor - Save code here

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

```
> library(ggplot2)
Loading required package: reshape
Loading required package: plyr
Attaching package: 'reshape'

The following object(s) are masked from 'package:plyr':
  rename, round_any

Loading required package: grid
Loading required package: proto
> qplot(displ, cyl, data = mpg)
> qplot(displ, hwy, data = mpg)
> 
```
Shortcuts

Learn them!

**In editor**
Cmd/ctrl + enter – send code to console
ctrl + 2 – move cursor to console

**In console**
Up arrow – retrieve previous command
ctrl + up arrow – search commands
ctrl + 1 – move cursor to editor
install.packages("ggplot2")
library(ggplot2)

?mpg
head(mpg)
str(mpg)
summary(mpg)

qplot(displ, hwy, data = mpg)
install.packages("ggplot2")
library(ggplot2)

?mpg
head(mpg)
str(mpg)
summary(mpg)

qplot(displ, hwy, data = mpg)
Scatter Plot Basics

```r
qplot(displ, hwy, data = mpg)
```
Additional Dimensions?

```
qplot(displ, hwy, colour=class, data=mpg)
```
Additional Dimensions?

\texttt{qplot(displ, hwy, colour=class, data=mpg)}

Legend chosen and displayed automatically.
Your Turn

Experiment with color, size, and shape aesthetics.

What's the difference between discrete or continuous variables?

What happens when you combine multiple aesthetics?
<table>
<thead>
<tr>
<th></th>
<th>Discrete</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Color</strong></td>
<td><img src="image" alt="Discrete color steps" /></td>
<td><img src="image" alt="Continuous color gradient" /></td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>discrete size steps</td>
<td>Linear mapping between radius and value</td>
</tr>
<tr>
<td><strong>Shape</strong></td>
<td><img src="image" alt="Discrete shape symbols" /></td>
<td>?</td>
</tr>
</tbody>
</table>
Faceting
= Small Multiples

Your turn!

qplot(displ, hwy, data=mpg) + facet_grid(. ~ cyl)
qplot(displ, hwy, data=mpg) + facet_grid(drv ~ .)
qplot(displ, hwy, data=mpg) + facet_grid(drv ~ cyl)
qplot(displ, hwy, data=mpg) + facet_wrap(~ class)
Faceting
= Small Multiples

Your turn!

qplot(displ, hwy, data=mpg) + facet_grid(. ~ cyl)
qplot(displ, hwy, data=mpg) + facet_grid(drv ~ .)
qplot(displ, hwy, data=mpg) + facet_grid(drv ~ cyl)
qplot(displ, hwy, data=mpg) + facet_wrap(~ class)

Summary

facet_grid() 2d grid, rows ~ cols, . for no split
facet_wrap() 1d ribbon wrapped into 2d
What's the problem here?

```
qplot(cty, hwy, data = mpg)
```
What's the problem here?

\texttt{qplot(cty, hwy, data = mpg, geom = "jitter")}
What's the problem here?

```r
qplot(cty, hwy, data = mpg, geom = "jitter")
```
How can we improve this plot?

```r
qplot(class, hwy, data = mpg)
```
How can we improve this plot?

```r
qplot(reorder(class, hwy), hwy, data = mpg)
```
How can we improve this plot?

```r
ggplot(reorder(class, hwy), hwy, data = mpg, geom = "boxplot")
```
How can we improve this plot?

```r
qplot(reorder(class, hwy), hwy, data=mpg, geom=c("jitter", "boxplot"))
```
Your Turn

Read the help for reorder. Redraw the previous plots with class ordered by **median hwy**.

How would you put the jittered points on top of the boxplots?
Diamonds

A bigger data set
Diamonds data

~54,000 round diamonds from http://www.diamondse.info

Carat, colour, clarity, cut

Total depth, table, depth, width, height

Price
Metrics of a diamond

\[ \text{depth} = \frac{z}{\text{diameter}} \]

\[ \text{table} = \frac{\text{table width}}{x} \times 100 \]
Your Turn

Inspect the data set
Barcharts vs Histograms

Nominal/categorical variables

Continuous variables
Let’s plot

With only one variable, qplot guesses that you want a bar chart or histogram

```r
qplot(cut, data = diamonds)
qplot(carat, data = diamonds)

# Change binwidth
qplot(carat, data = diamonds, binwidth = 1)
qplot(carat, data = diamonds, binwidth = 0.1)
qplot(carat, data = diamonds, binwidth = 0.01)

last_plot() + xlim(0, 3)

resolution(diamonds$carat)
```
Always experiment with the bin width!
qplot(depth, data = diamonds, binwidth = 0.2) + xlim(55, 70)
Additional Dimensions

\texttt{qplot(depth, data = diamonds, binwidth = 0.2, fill = cut) + xlim(55, 70)}
Additional Dimensions

qplot(depth, data = diamonds, binwidth = 0.2, fill = cut) + xlim(55, 70) + facet_wrap(~ cut)
Your Turn

Explore the distribution of price. What is a good binwidth to use? 

Hint: Diamonds are expensive!

Practice zooming in on regions of interest.

How does price vary with color, cut, or clarity?
Frequency Histogram

They’re helpful, but come with caveats

\texttt{qplot}(\texttt{price}, \texttt{data = diamonds}, \texttt{binwidth = 500}) + \texttt{facet_wrap(\sim \texttt{cut})}
Frequency Histogram
They’re helpful, but come with caveats

What makes it difficult to compare the shape of the distributions?

qplot(price, data = diamonds, binwidth = 500) + facet_wrap(~ cut)
Frequency Histogram
They’re helpful, but come with caveats

# Large distances make comparisons hard
qplot(price, data = diamonds, binwidth = 500) +
facet_wrap(~ cut)

# Stacked heights hard to compare
qplot(price, data = diamonds, binwidth = 500, fill = cut)

# Much better - but still have differing relative abundance
qplot(price, data = diamonds, binwidth = 500,
       geom = "freqpoly", colour = cut)

# Instead of displaying count on y-axis, display density
# .. indicates that variable isn't in original data
qplot(price, ..density.., data = diamonds, binwidth = 500,
      geom = "freqpoly", colour = cut)
Density Histogram
Shows relative distribution better

```r
qplot(price, ..density.., data = diamonds, binwidth = 500, geom = "freqpoly", colour = cut)
```
Where Next?
Learn more about

- Aggregating your data: plyr
- Working with dates: lubridate
- Regular expressions: stringr
- A consistent philosophy of data: google “tidy data”
- ggplot2: [http://blog.ggplot2.org/](http://blog.ggplot2.org/) + ggplot2 mailing list
Other Resources

- The art of R programming http://amzn.com/1593273843
- Data manipulation with R http://amzn.com/0387747303
- http://www.r-bloggers.com/
- http://stackoverflow.com/questions/ tagged/r
This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 United States License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/3.0/us/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.
Next Lecture
Dashboards. Guest lecture by Stephen Few