Before we start…

Please download/install the latest R version and RStudio at:

http://www.rstudio.com

http://cran.r-project.org

Make sure you have a GitHub.com account.

The slides will be made available online after the course

July, 2014 · ulrich atz · @statshero

Introductions

- Name
- Role/department
- Your experience of R to date
- What you'd hope to gain by the end of today / learning from yesterday?

Agenda - Today

1. Intro to R
2. Setting up a data analysis project
3. *** Lunch ***
4. Visualisations
5. Putting it all together

Introductions to R

Why use R?

Open source / free

Widely used (> 2 million and growing)

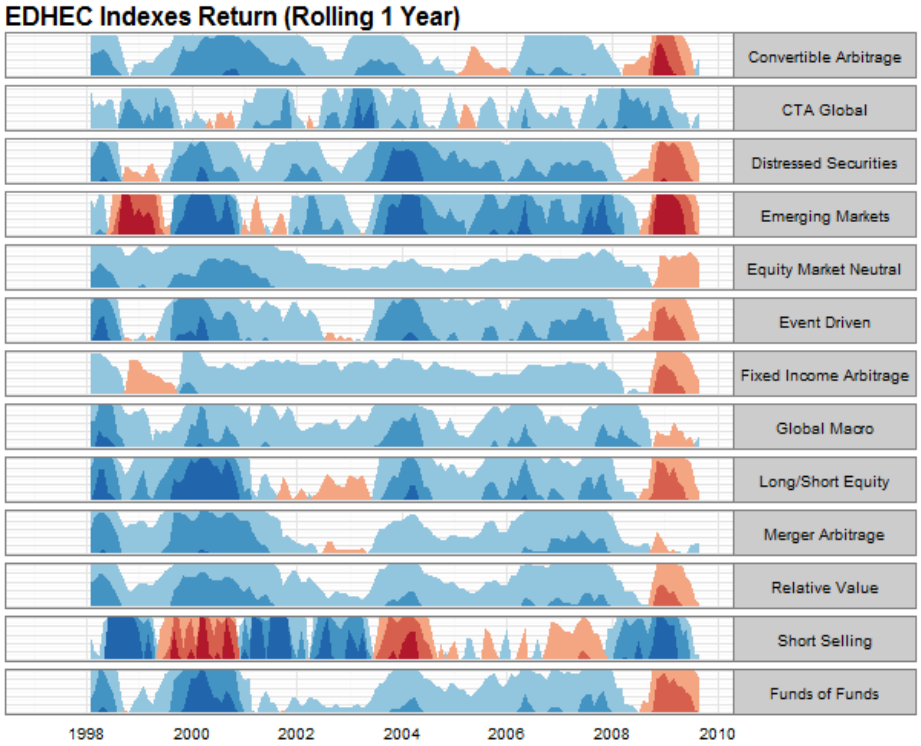
R has an incredible community (> 6,000 packages)

Used by Facebook, Google (apparently 500+ users), weather forecasts, finance industry

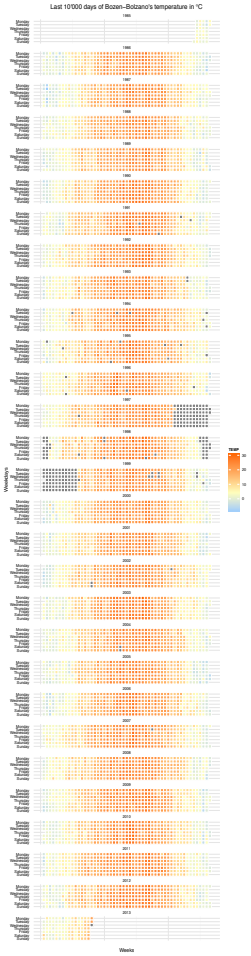
*During 2013 alone, R added more functions than SAS Institute has written in its entire history!**

*http://r4stats.com/articles/popularity

What R graphics can do

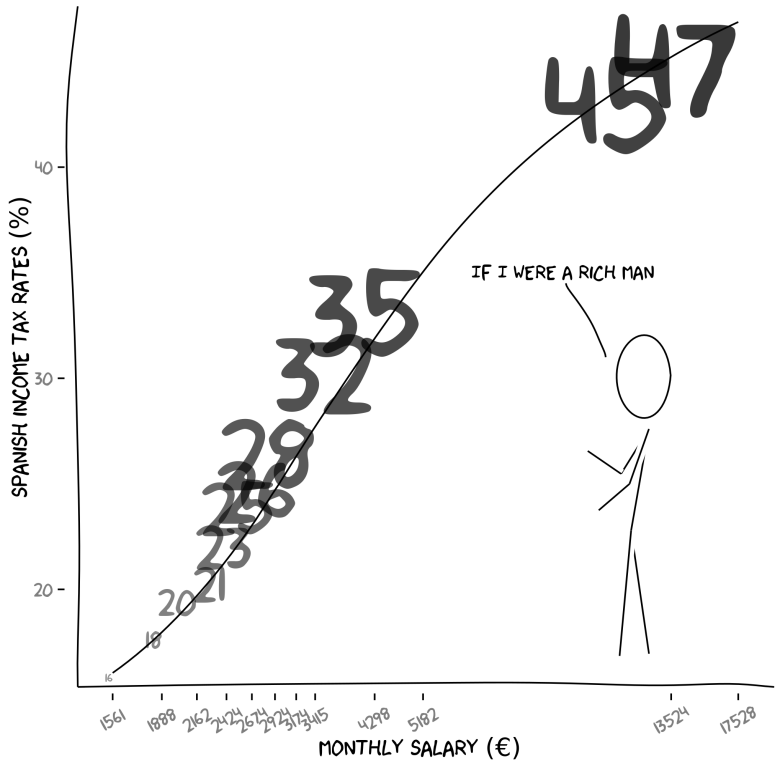


http://timelyportfolio.blogspot.co.uk/2012/08/horizon-on-ggplot2.html



What R graphics can do (2)

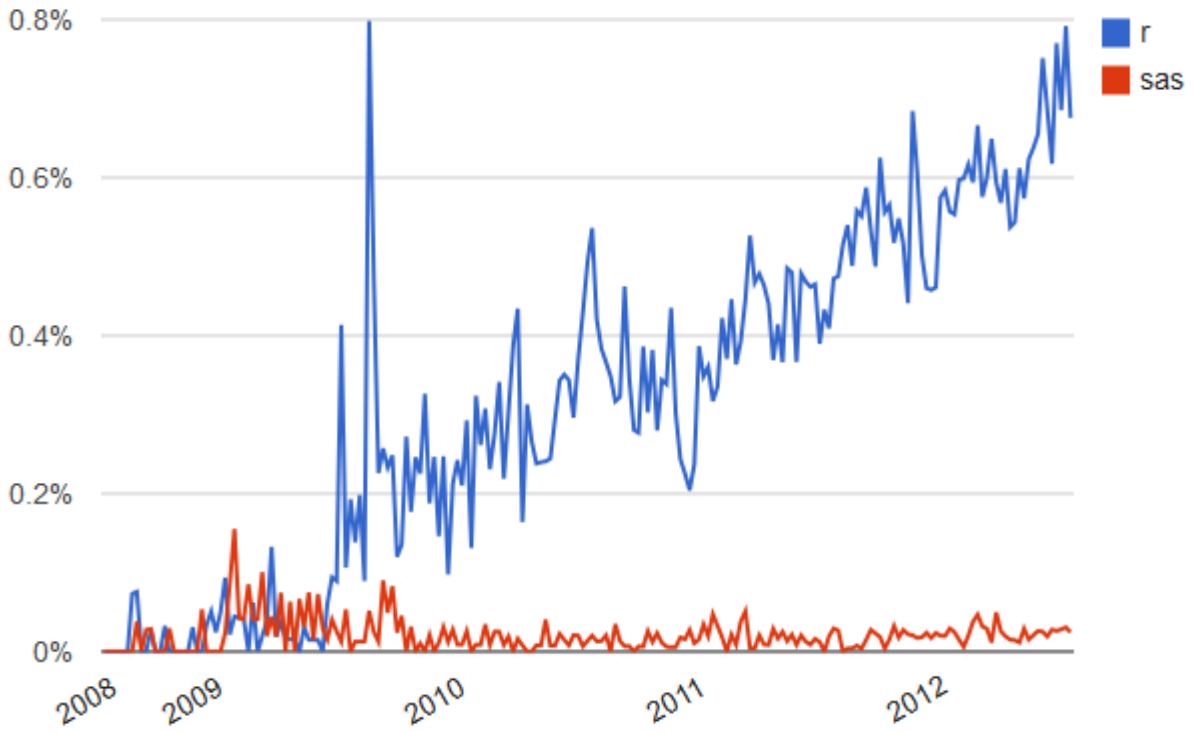
What R graphics can do (3)



http://xkcd.r-forge.r-project.org/

vignette("xkcd-intro")

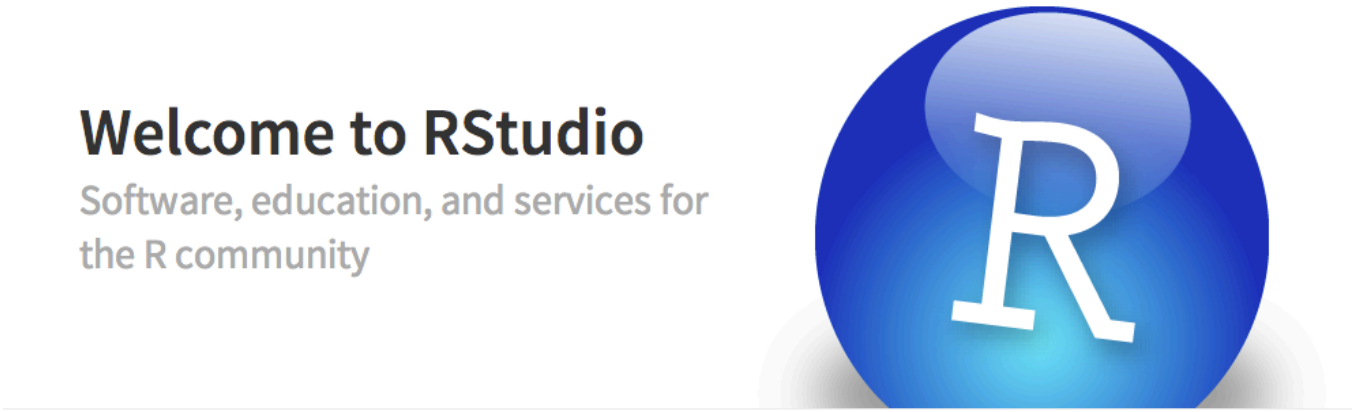
Posts on stackoverflow



http://r4stats.com/articles/popularity/

RStudio

http://www.rstudio.com/



Getting started

Spend five minutes to get familiar with the RStudio interface. Some things to try:

1. Understand the purpose of the four windows
2. Have a look at `help.start()`
3. Bonus: go to settings and customise

Popular packages

Base R is extended by the community.

Anyone can write one!

Top Ranked CRAN Packages

Week | Month | **All time**

#		Package	#⬇
1	-	ggplot2	677993
2	-	plyr	578586
3	-	digest	530768
4	-	stringr	530631
5	-	colorspace	482884

Getting help

- Use the help() function, shortcut is `?plot`
- Type your question into google
- Ask on the email list or stackoverflow
- http://www.rseek.org/

RStudio – keyboard shortcuts

The three most important keyboard shortcuts:

1. **Tab** is a generic auto-complete function.
2. **Control + the up arrow** (command + up arrow on a Mac) is a similar auto-complete tool. (This works only in the interactive console, not in the code editor window.)
3. **Control + enter** (command + enter on a Mac) takes the current line of code in the editor, sends it to the console and executes it. If you select multiple lines of code in the editor and then hit ctrl/cmd + enter, all of them will run.

http://www.computerworld.com/s/article/9239625/Beginner_s_guide_to_R_Introduction

Explore a dataset with R

- Use classic Fair (1978) dataset on extramarital affairs
- Import with `read.csv()`
- http://bit.ly/BIS-links

Variable description

Description of Variable	Values of Variable	Mean Value
	PT Tape	
How often engaged in extramarital sexual intercourse during the past year	0 = none, 1 = once, 2 = twice, 3 = 3 times, 7 = 4–10 times, 12 = monthly, 12 = weekly, 12 = daily	1.46
Sex	0 = female, 1 = male	.476
Age	17.5 = under 20, 22.0 = 20–24, 27.0 = 25–29, 32.0 = 30–34, 37.0 = 35–39, 42.0 = 40–44, 47.0 = 45–49, 52.0 =	32.5

Variable description

	3 = slightly, 2 = not at all, 1 = anti	
Level of education	9.0 = grade school, 12.0 = high school graduate, 14.0 = some college, 16.0 = college graduate, 17.0 = some graduate work, 18.0 = master's degree, 20.0 = Ph.D., M.D., or other advanced degree	16.2
Occupation	1–7, according to Hollingshead classification (reverse numbering)	4.19
How rate marriage	5 = very happy, 4 = happier than average, 3 = average, 2 = somewhat unhappy, 1 = very unhappy	3.93

Starting point for interactive work

- `str()`
- `names()`
- `summary()`
- `head()`
- `tail()`
- `table()`

VERSION CONTROL?

Have you ever…?

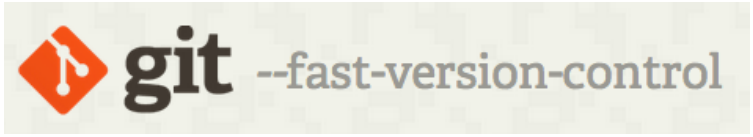
- Made a mistake and wanted to revert back?
- Lost code or had a backup that was too old?
- Wanted to play around with different ideas?
- Wanted to see the difference between two (or more) versions of your code?
- Wanted to suggest a change to someone else's analysis?
- Wanted to share your analysis, or let other people work on it?
- Wanted to experiment with a new feature without interfering with working code?

R and version control for the solo data analyst

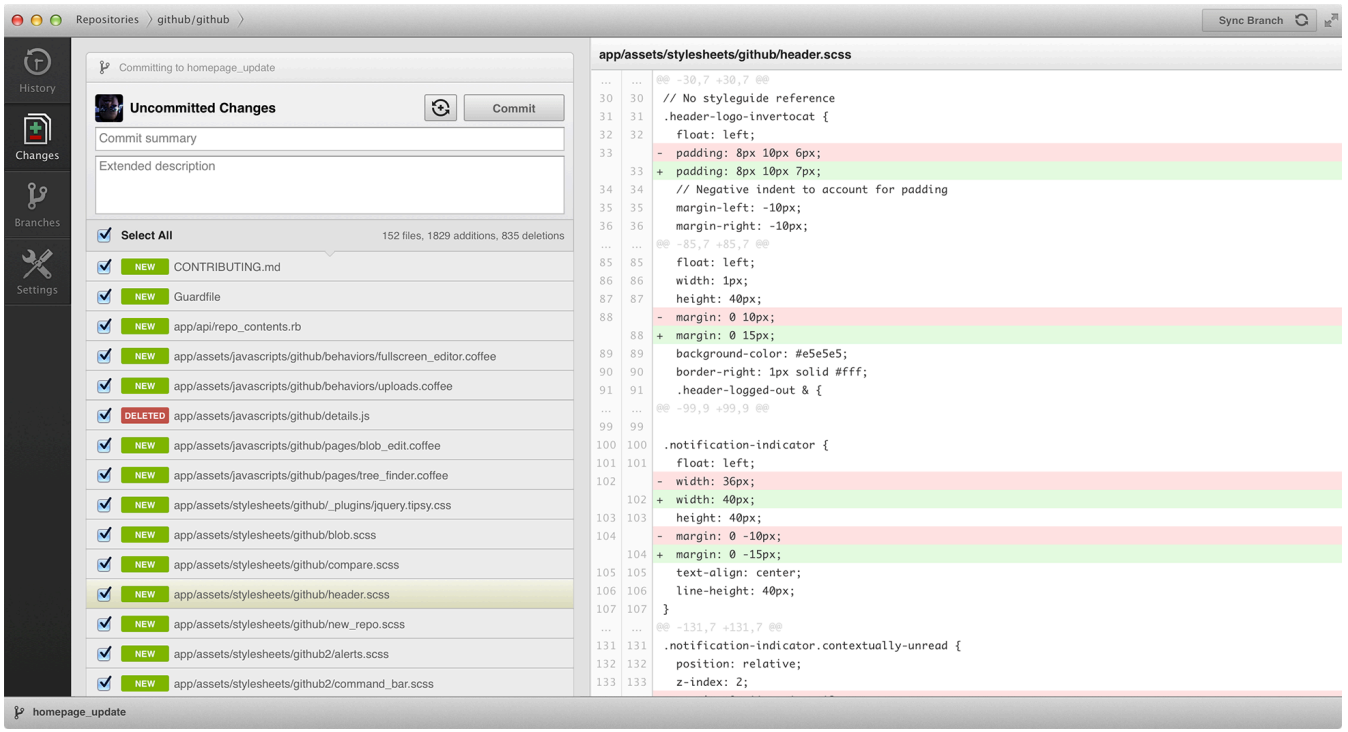
- Version control ≠ backup
- Can be lightweight
- Potentially reduces complexity, e.g. file management

http://stackoverflow.com/questions/2712421/r-and-version-control-for-the-solo-data-analyst

Different software, different ways



GitHub for Mac/Windows



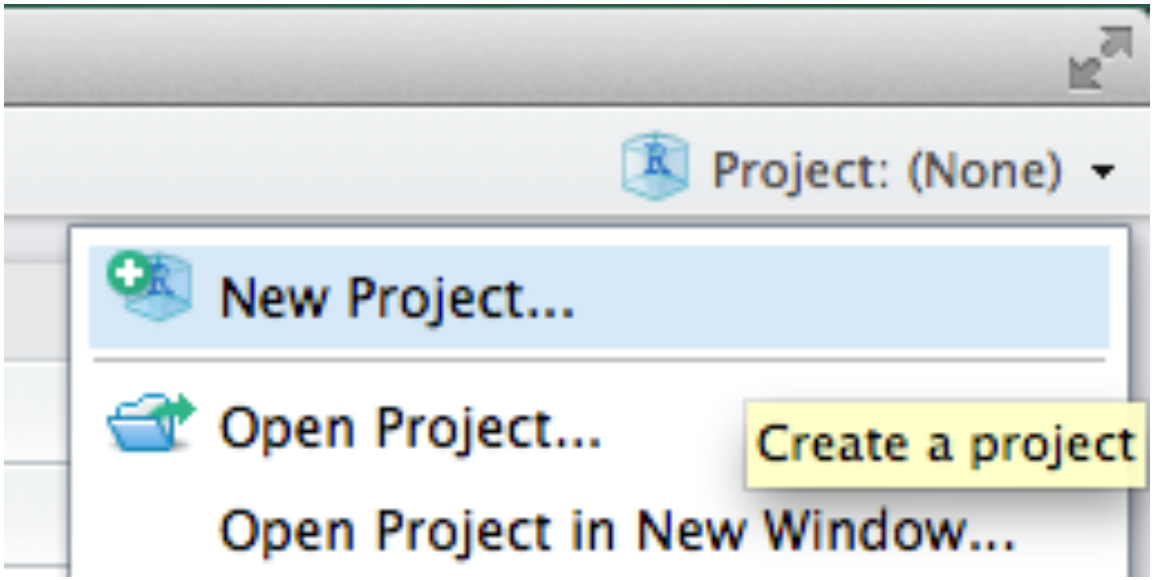
Before we use Git

Create a working directory (= “repository”)

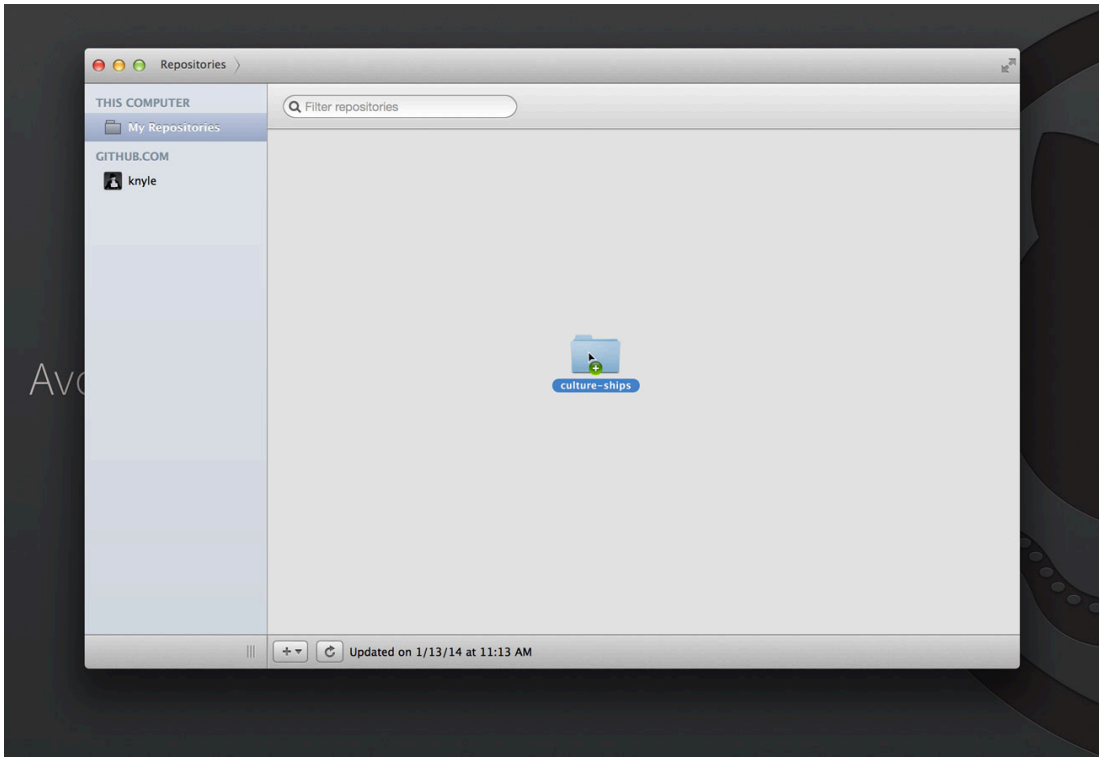
Save your files there.

(It usually includes a README file, too.)

New project in RStudio

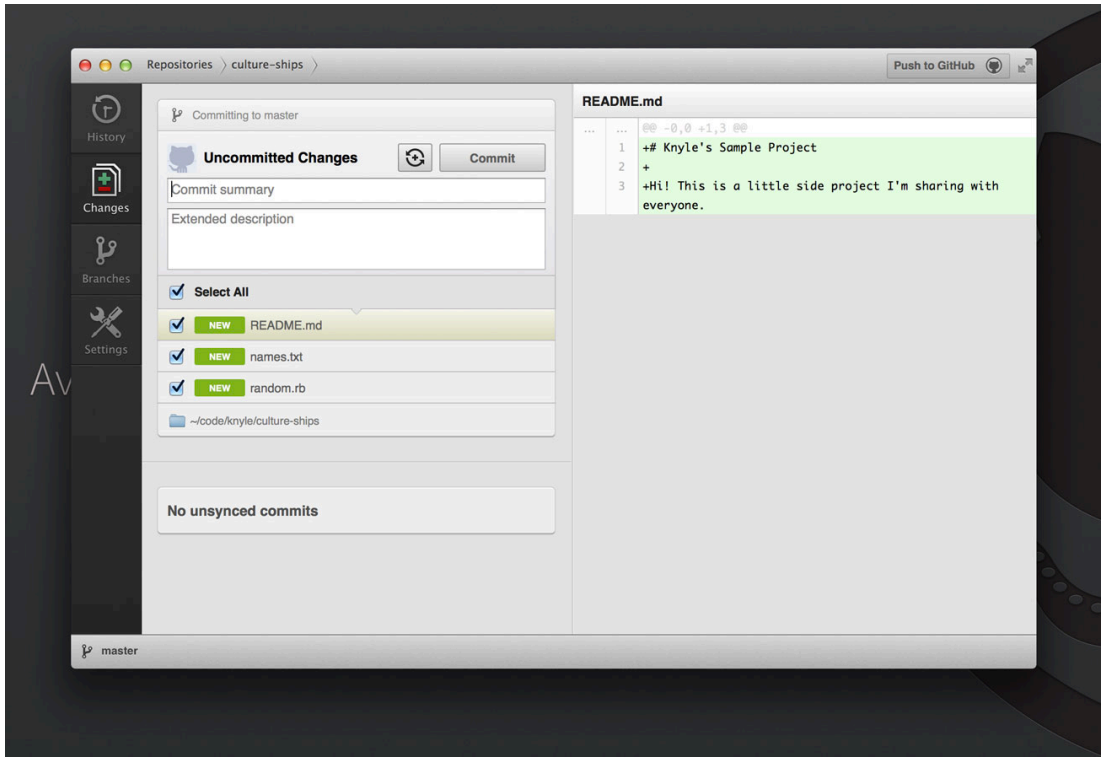


Getting it on GitHub



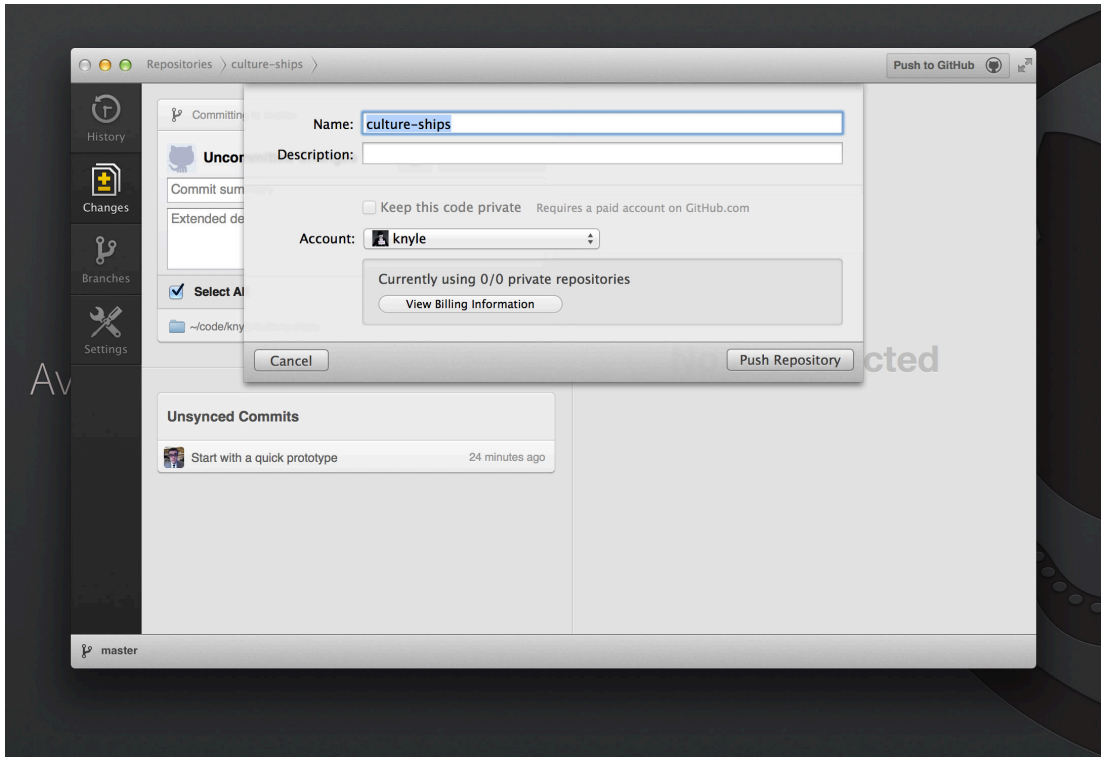
https://guides.github.com/introduction/desktop/

Your first commit



https://guides.github.com/introduction/desktop/

Push your code to GitHub.com



https://guides.github.com/introduction/desktop/

One more commit for good measure

http://xkcd.com/1296

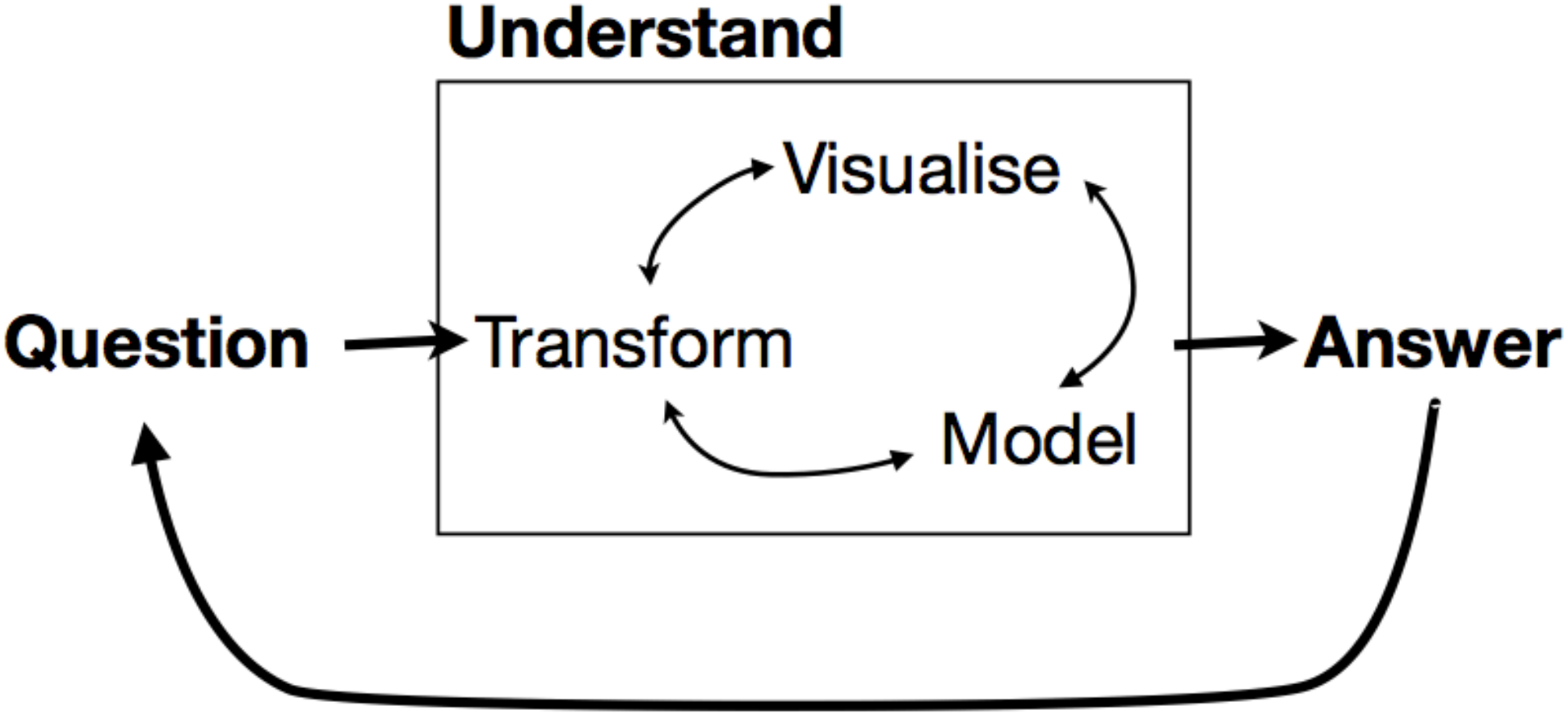
Clone my repository

https://github.com/theodi/BIS-day-2

Run the R script.

Bonus: make a branch and a pull request

WHY ARE WE DOING THIS?



Source: Visualisation in R with ggplot2, Garrett Grolemund and Winston Chang

A Model for Data Analysis

1.1 **FIND** reliable data sources

1.2 Understand its relevance and **LICENCE**

1.3 Visualise and **UNDERSTAND** your data

2.1 **CLEAN** your data

2.2 **TRANSFORM** it where useful

2.3 **COMBINE** it with other data sets

3.1 **REDUCE** and find the story

3.2 Think and understand the **CONTEXT**

3.3 Do your results pass a **SENSE-CHECK**?

Open data from e.g. data.gov.uk

BASIC GRAPHICS IN R

Graphics in R

1. base
2. qplot
3. ggplot

R offers three main types of graphics:

1. traditional from the base installation
2. the lattice package
3. ggplot2

Comparison of three R graphics packages

1. base
2. qplot
3. ggplot

	Traditional	lattice	ggplot2
Automatic output for different objects	Yes	No	Yes
Automatic legends	No	Sometimes	Yes
Easily repeats plots for different groups	No	Yes	Yes
Easy to use with multiple data sources	Yes	No	Yes
Consistent functions	No	No	Yes
Attractiveness of default settings	Fair	Fair	Excellent
Underlying graphics system	Traditional	Grid	Grid

Adapted from: R. Muenchen, R for SAS and SPSS users

Basic plots in R

1. base
2. qplot
3. ggplot

- `?plot`
- `methods(plot)`
- `library(help = "datasets”)`
- `data(yourDataset)`

> Pick one and plot some variables

Find examples for the following plots + try at least one option

1. base
2. qplot
3. ggplot

- `hist()`
- `qqnorm()`
- `boxplot()`
- `smoothScatter()`
- `dotchart()`
- `barplot()`

- `rug()`
- `plot(table())`
- `plot(density())`

- Use `data(iris)` or the help example if your dataset is not suited.

Exploring the Fair data set

1. base
2. qplot
3. ggplot

- What problems can you identity with visualising this data set?
- How can we solve the issue of overplotting categorical variables?
- What are “quick insights” you can present to the group?

THE BASICS OF GGPLOT2

Yield data from a Minnesota barley trial

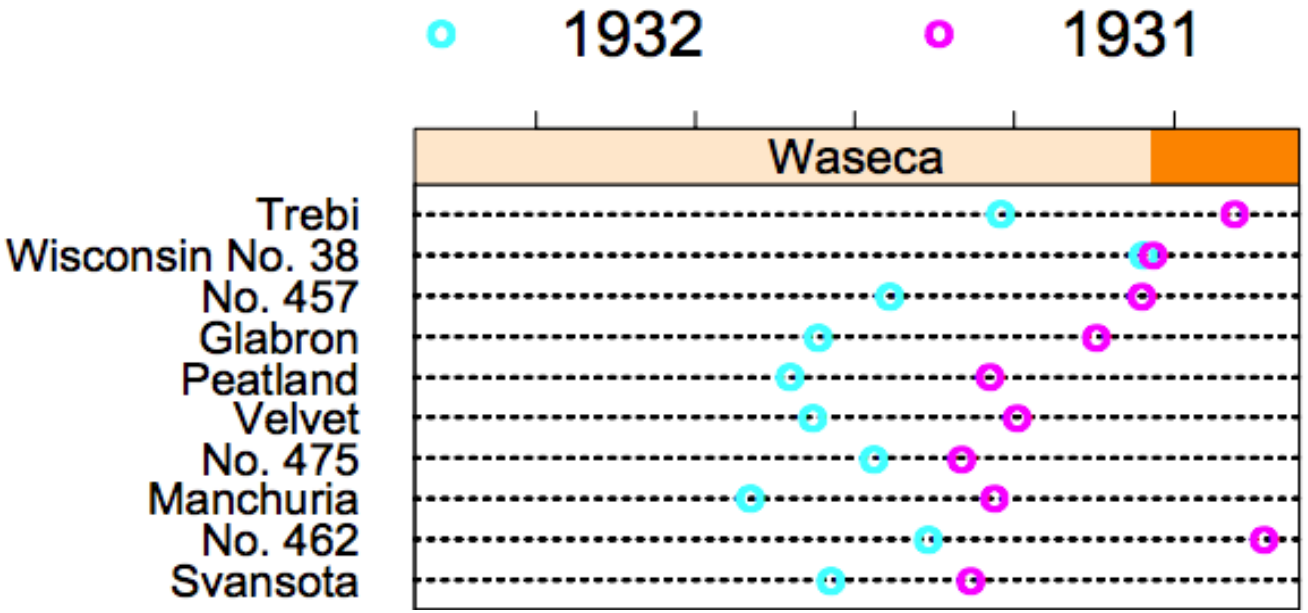
1. base
2. qplot
3. ggplot

```
library(ggplot2)
library(lattice)

# Look at the data we're going to use
?barley
head(barley)
str(barley)
```

Yield data from a Minnesota barley trial

1. base
2. qplot
3. ggplot



http://cm.bell-labs.com/cm/ms/departments/sia/doc/trellis.user.pdf

Bar charts and histograms

- `qplot(x, data)` – it will try to guess the best graph depending on what you supply.
- Experiment with the binwidth option!
- `resolution()` is a nice trick.
- `last_plot()` another one.

Quick plot (qplot)

1. base
2. qplot
3. ggplot

- `qplot(x, y, data=yourdata)`
- Try some options
- `xlab, ylab`

Additional variables

1. base
2. qplot
3. ggplot

- We can display additional variables with **aesthetics** (like shape, colour, size).
- Experiment with color, size, and shape aesthetics.
- What's the difference between discrete or continuous variables?

Aesthetics

1. base
2. qplot
3. ggplot

	Discrete	Continuous
Color	Rainbow of colors	Gradient from light blue to dark blue
Size	Discrete size steps	Linear mapping between radius and value
Shape	Different shape for each	Shouldn't work

Garrett Grolemund, 2012, ggplot2 basics.

Reordering and boxplots

1. base
2. qplot
3. ggplot

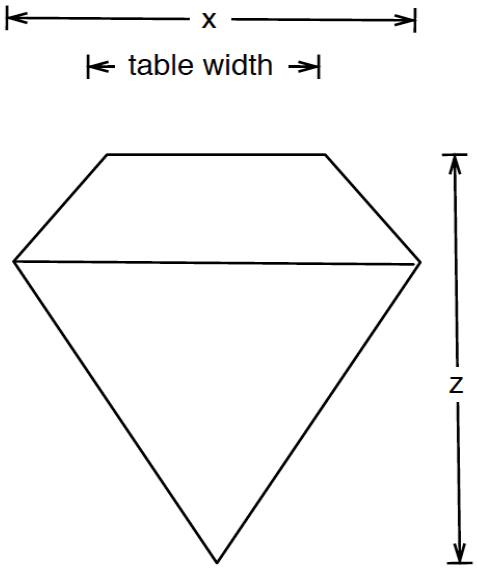
- Let's use the `reorder()` function.
- What does `geom="jitter"` do?
- What does `geom="boxplot"` do?
- What happens if you combine both of them?

DIAMONDS

Diamonds data

- Standard data set from the `ggplot2` package
- ~54,000 diamonds
- Carat, colour, clarity, cut
- Total depth, table, depth,
- Width, height
- Price

Diamonds

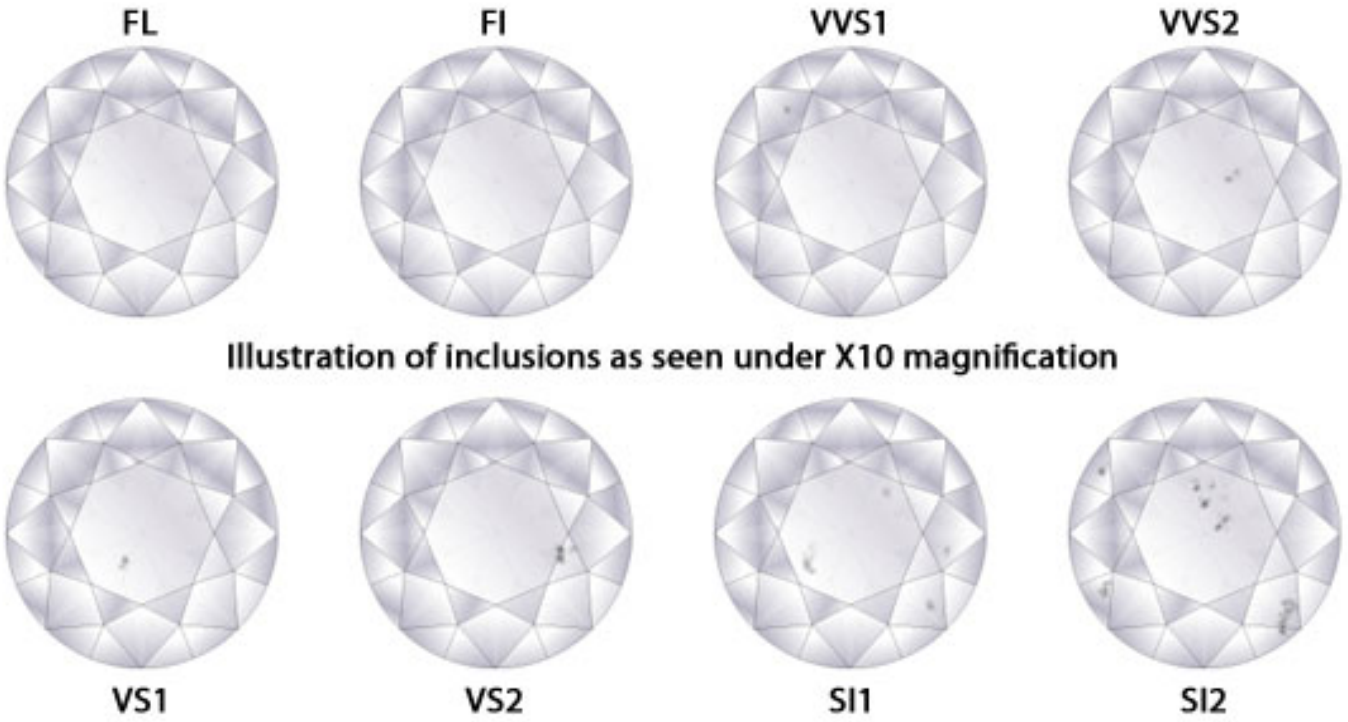


depth = z / diameter
table = table width / x * 100

Colors

http://www.mlcld.com/colour.aspx

Clarity



http://www.thediamondsexperts.com/diamonds-guide

Explore the Diamonds data

Use the qplot function and anything that you find useful.

For example:

```
head()
str()
```

Big scatterplots

- `qplot(carat, price, data=diamonds)`

- What are solutions to the overplotting problem?
- Brainstorm for 2 minutes.

Ideas

Idea	ggplot
Small points	size = I(0.25)
Transparency	alpha = I(1/50)
Jittering	geom = "jitter"
Smooth curve	geom = "smooth"
2d bins	geom = "bin2d" or geom = "hex"
Density contours	geom = "density2d"

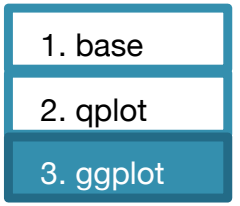
Garrett Grolemund, 2012, ggplot2 basics.

Let’s go again

```
qplot(x, y, data = diamonds)
# Very basic cleaning
diamonds$x.o <- diamonds$x
diamonds$y.o <- diamonds$y
diamonds$x[diamonds$x == 0] <- NA
diamonds$y[diamonds$y == 0] <- NA
diamonds$y[diamonds$y > 12] <- NA
```

ADVANCED GGPLOT2

The grammar of graphics



- The ggplot2 package is based on the same concepts as SPSS's GPL.
- qplot is the shortcut for the ggplot logic.

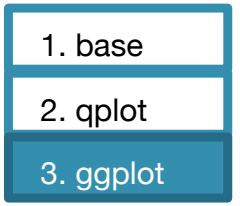
"If any number of magnitudes are each the same multiple of the same number of other magnitudes, then the sum is that multiple of the sum."

Euclid, ~300 BC

$$m(\Sigma x) = \Sigma(mx)$$

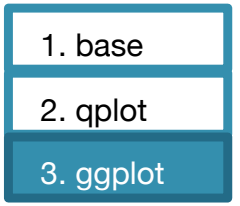
http://www.slideshare.net/hadley/grammar-of-graphics-past-present-future

What is a layer?



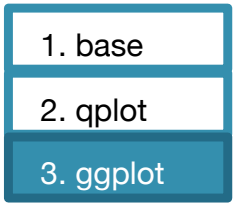
- Data
- Mappings from variables to aesthetics (aes)
- A geometric object (geom)
- A statistical transformation (stat)
- A position adjustment (position)

```
layer(data, mapping, geom, stat,
position, ...)
```



- layer(
  data = diamonds,
  mapping = aes(x = carat),
  geom = "bar",
  stat = "bin",
  position = "stack"
)

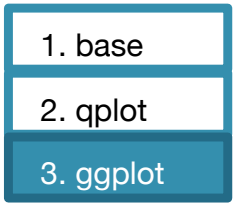
Creating the plot




```
ggplot() + geom_histogram(aes(carat),
    data = diamonds)


p <- ggplot() +
geom_histogram(aes(carat), data=diamonds)
class(p)
```

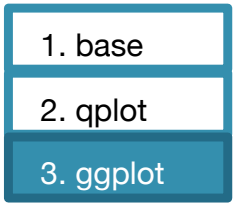
In practice




```
 # Multiple layers
ggplot() +
  geom_point(aes(carat, price), data = diamonds) +
  geom_smooth(aes(carat, price), data = diamonds)

# Avoid redundancy:
ggplot(diamonds, aes(carat, price) +
  geom_point() +
  geom_smooth()
```

Convert qplot to ggplot graphs



- `qplot(carat, price, data = diamonds)`

- `qplot(log10(carat), log10(price), data = diamonds, colour = color) + geom_smooth(method = "lm")`

Geometric objects

Name	Description
abline	Line, specified by slope and intercept
area	Area plots
bar	Bars, rectangles with bases on y-axis
blank	Blank, draws nothing
boxplot	Box-and-whisker plot
contour	Display contours of a 3d surface in 2d
crossbar	Hollow bar with middle indicated by horizontal line
density	Display a smooth density estimate
density_2d	Contours from a 2d density estimate
errorbar	Error bars
histogram	Histogram
hline	Line, horizontal
interval	Base for all interval (range) geoms
jitter	Points, jittered to reduce overplotting
line	Connect observations, in order of x value
linerange	An interval represented by a vertical line
path	Connect observations, in original order
point	Points, as for a scatterplot
pointrange	An interval represented by a vertical line, with a point in the middle
polygon	Polygon, a filled path
quantile	Add quantile lines from a quantile regression
ribbon	Ribbons, y range with continuous x values
rug	Marginal rug plots
segment	Single line segments
smooth	Add a smoothed condition mean
step	Connect observations by stairs
text	Textual annotations
tile	Tile plot as densely as possible, assuming that every tile is the same size
vline	Line, vertical

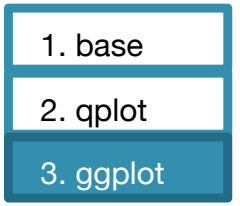
1. base
2. qplot
3. ggplot

Statistical transformations

1. base
2. qplot
3. ggplot

Name	Description
bin	Bin data
boxplot	Calculate components of box-and-whisker plot
contour	Contours of 3d data
density	Density estimation, 1d
density_2d	Density estimation, 2d
function	Superimpose a function
identity	Don't transform data
qq	Calculation for quantile-quantile plot
quantile	Continuous quantiles
smooth	Add a smoother
spoke	Convert angle and radius to xend and yend
step	Create stair steps
sum	Sum unique values. Useful for overplotting on scatterplots
summary	Summarise y values at every unique x
unique	Remove duplicates

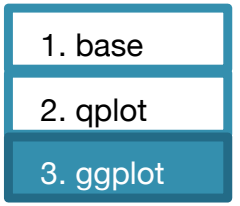
Exercise: Are diamonds symmetric?



Remember some cleaning is useful:

- `diamonds$x[diamonds$x == 0] <- NA`
- `diamonds$y[diamonds$y == 0] <- NA`
- `diamonds$y[diamonds$y > 12] <- NA`

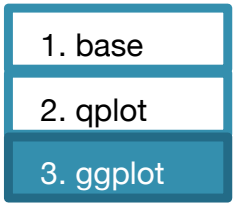
Exercise: Are diamonds symmetric?



- Define a sensible cut-off for outliers!
- Use ggplot() again to produce the scatterplot.

- Other ideas:
- `diamonds$area <- diamonds$x * diamonds$y`
- `diamonds$lratio <- log10(diamonds$x / diamonds$y)`

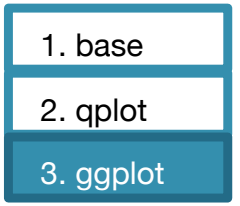
Example on a smaller dataset




```
sdiamonds <- diamonds[sample(nrow(diamonds),
1000), ]
sdiamonds$cut <- factor(sdiamonds$cut, levels =
c("Ideal", "Very Good", "Fair", "Good", "Premium"))

# Repeat first example with new order
p <- ggplot(sdiamonds, aes(carat, ..density..)) +
        geom_histogram(binwidth = 1)
p + facet_grid(. ~ cut)
```

facet_wrap




```
d <- ggplot(diamonds, aes(carat, price)) +
geom_hex()
d + facet_wrap(~ color)
```

What does `scales = "free_y"` do?


BY SA

Exercise: Save the Titanic

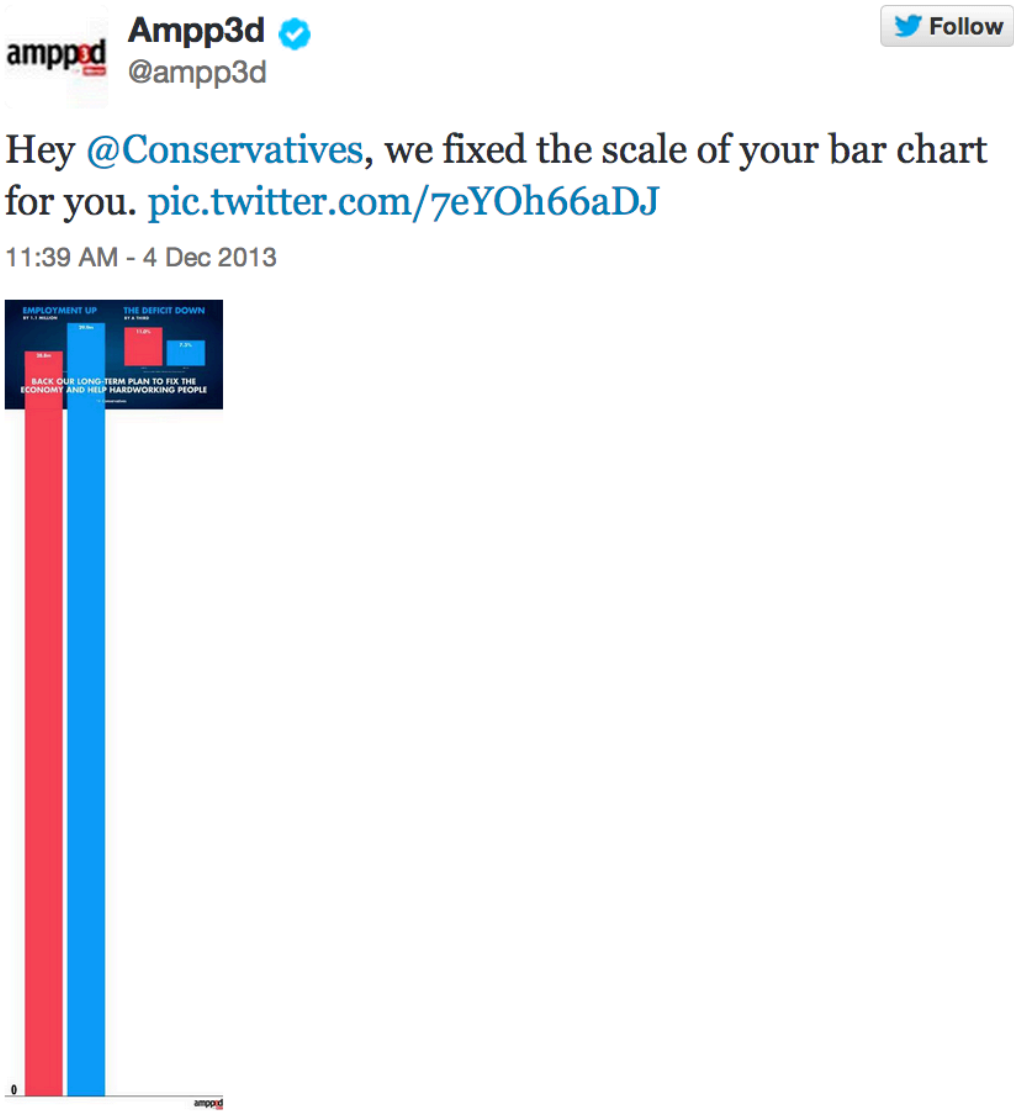
http://bit.ly/BIS-titanic

http://en.wikipedia.org/wiki/File:Stöwer_Titanic.jpg

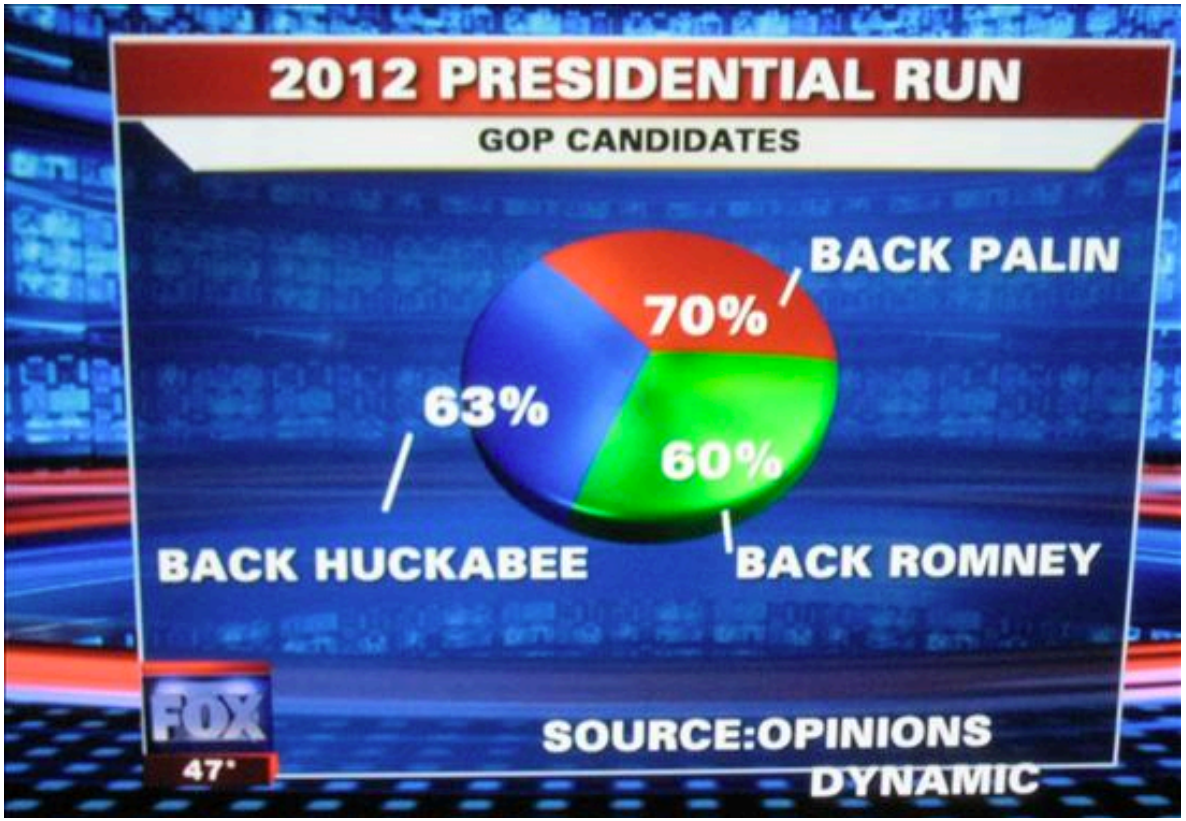
PRESENTING YOUR STORY

HOW (NOT) TO VISUALISE YOUR DATA

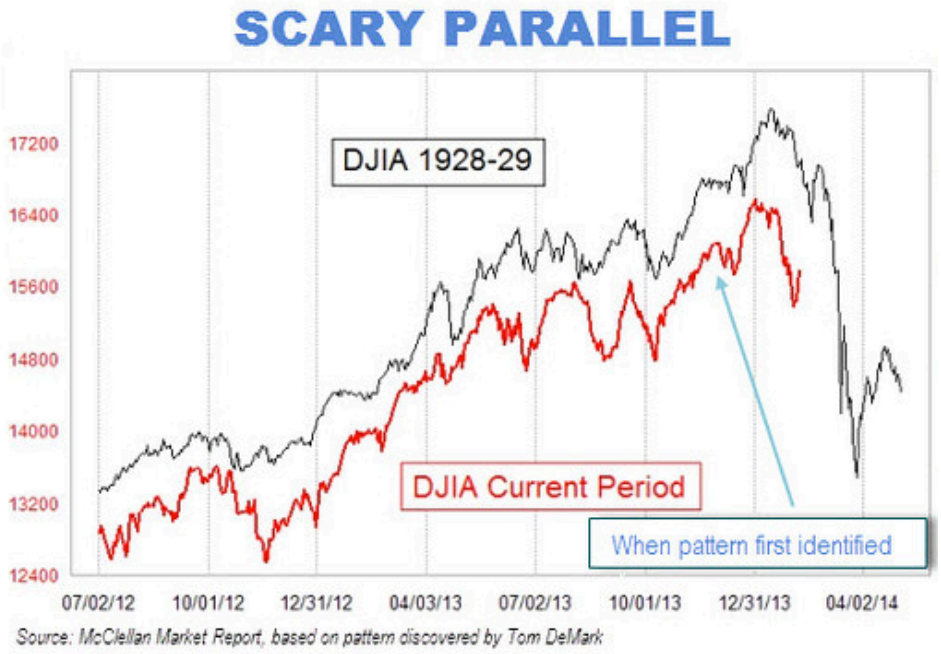
Y-axis does not start at zero

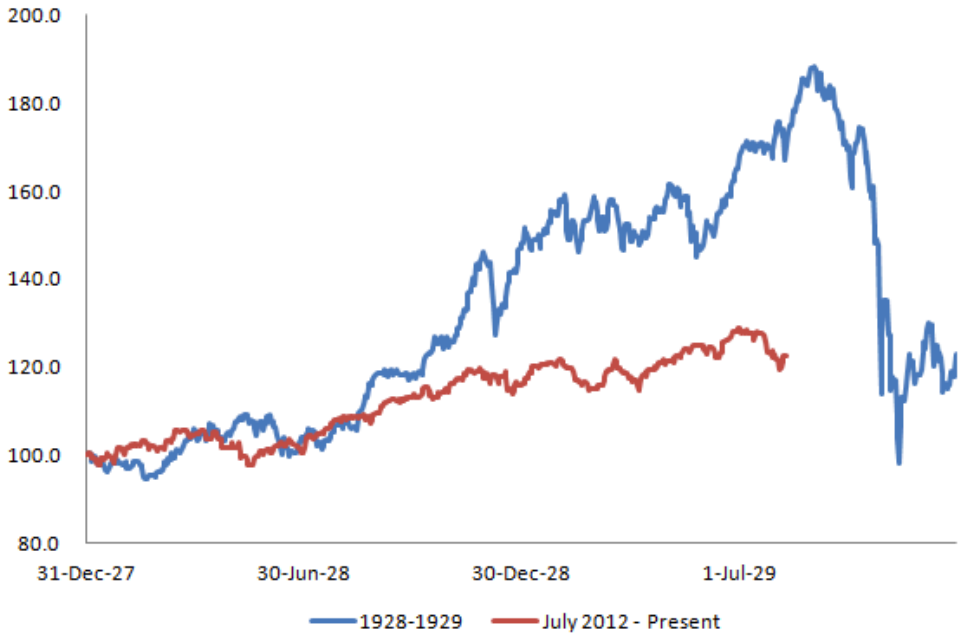


Non-traditional use of chart types



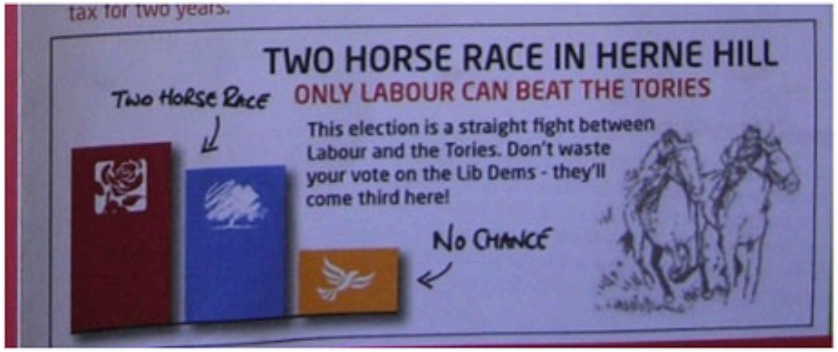
Arbitrariness of double axes



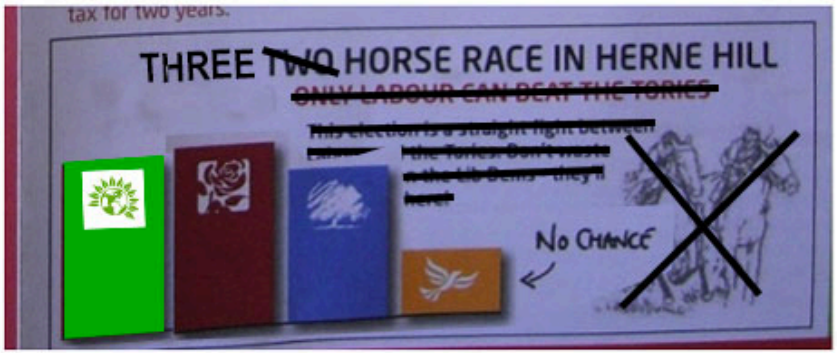


Selective use of data

WRONG

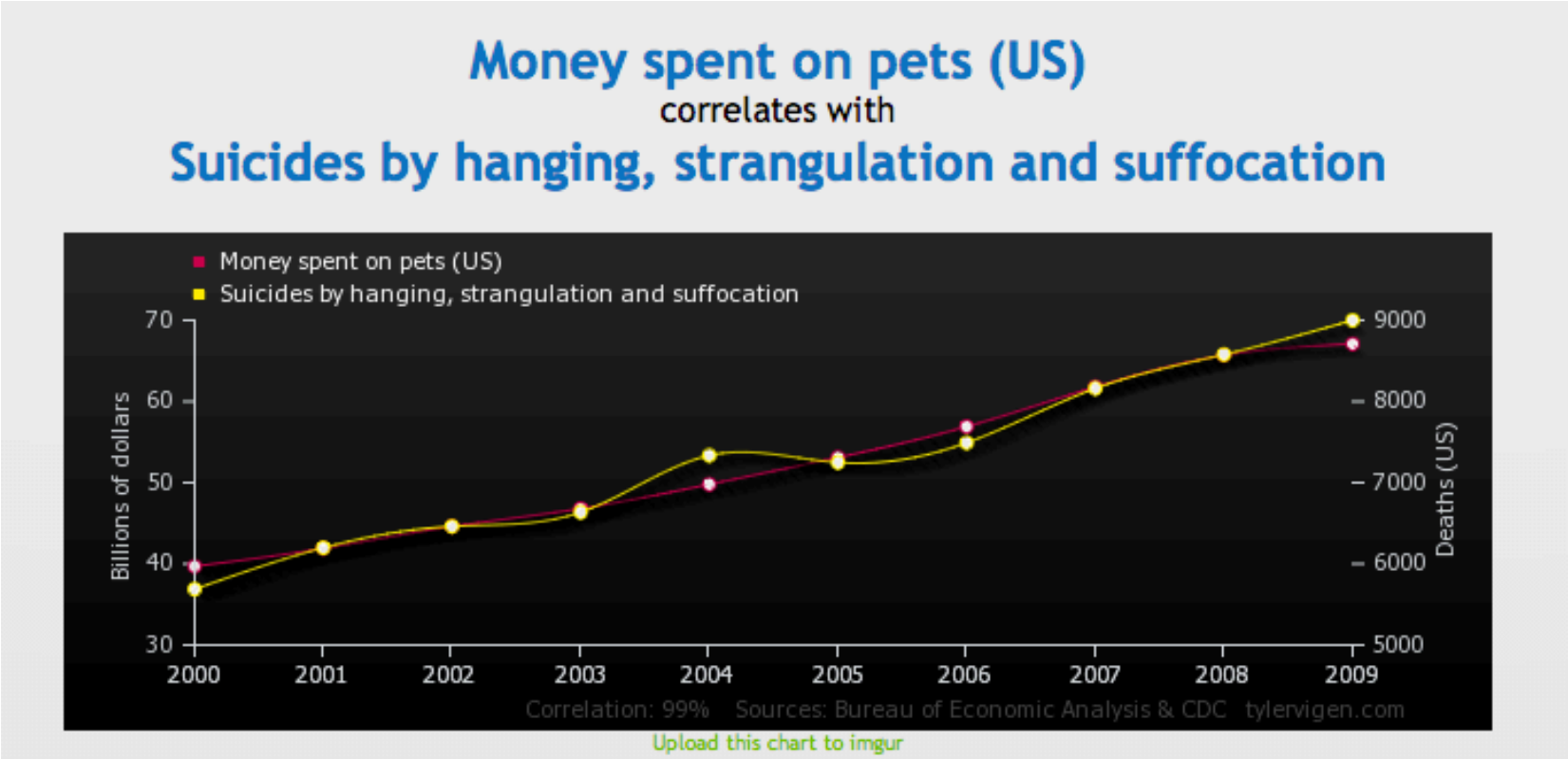


FIXED



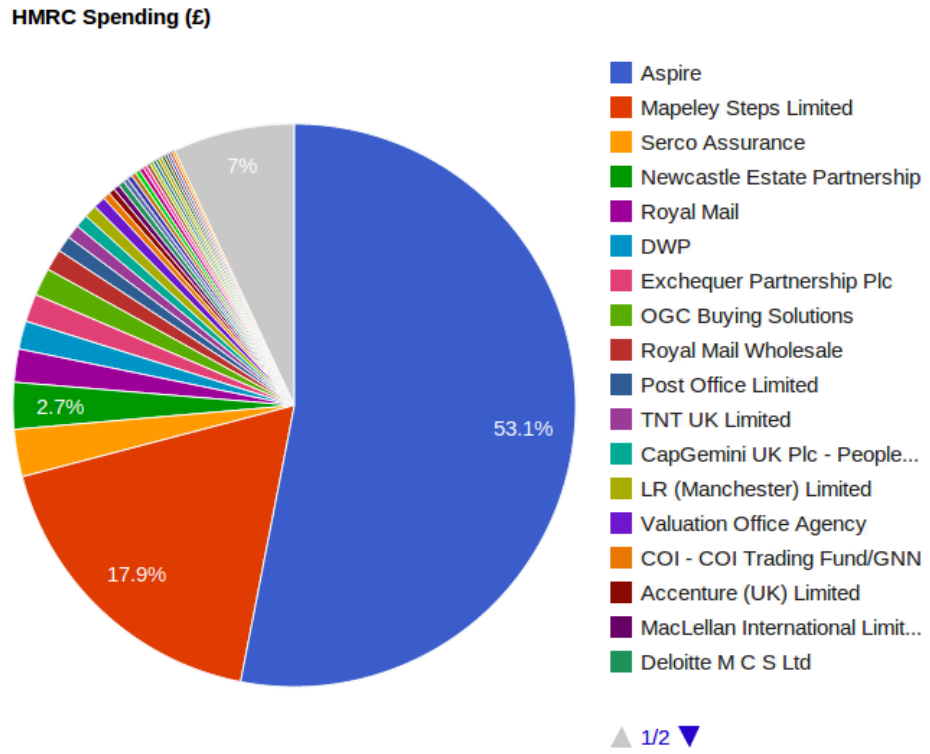
http://ampp3d.mirror.co.uk/2014/05/16/11-mistakes-that-will-drive-data-nerds-crazy/

Finding a meaningless correlation

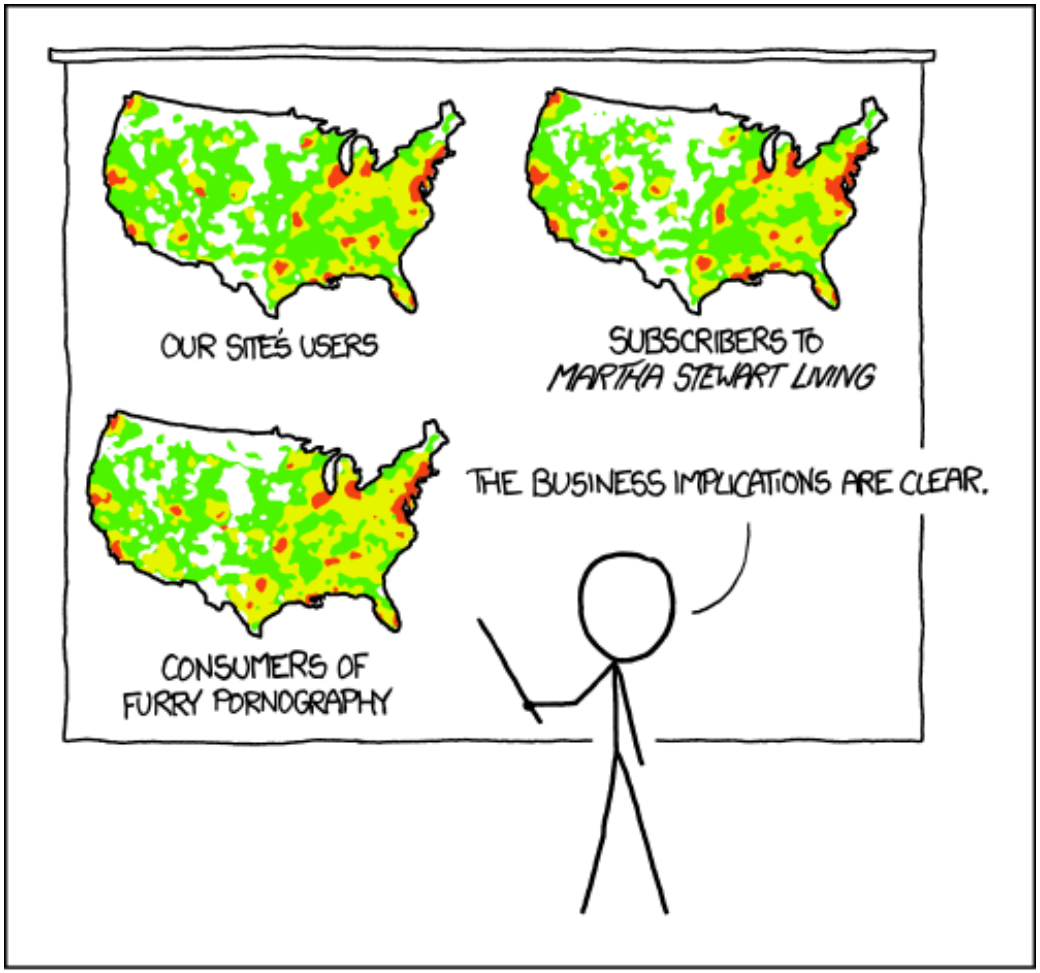


http://www.tylervigen.com/

Charts that use ALL THE DATA



Assuming maps are perfect



http://xkcd.com/1138/

CONCLUSIONS AND TIME FOR QUESTIONS

Feedback

Please complete our form:

http://bit.ly/odifeedback

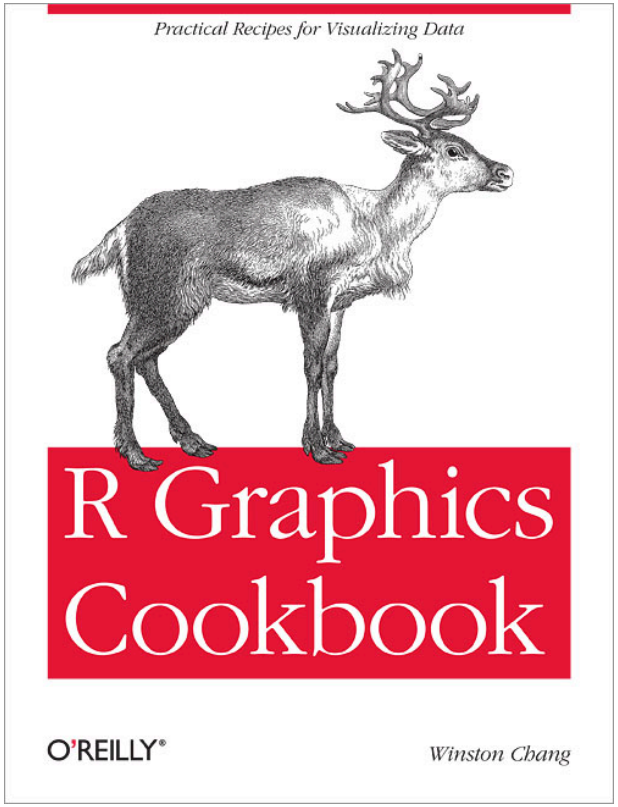
Thank you!

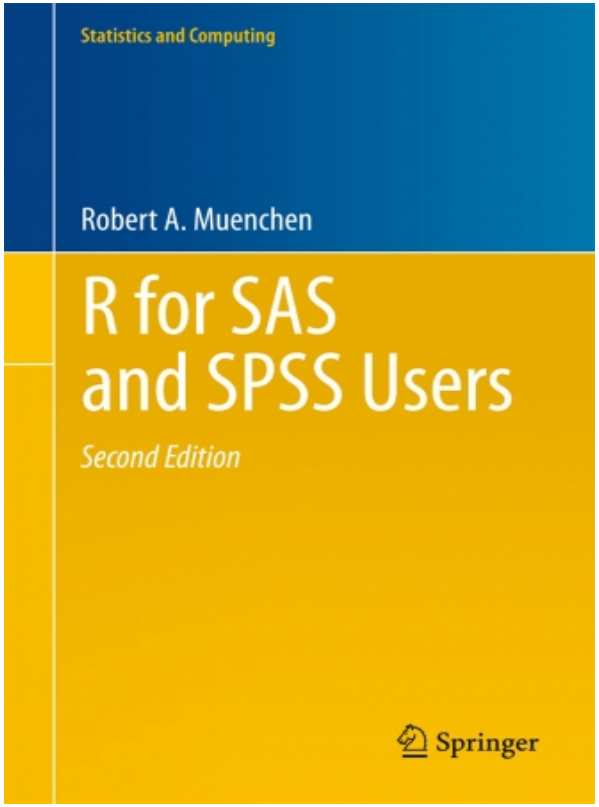
Further reading and links -
http://bit.ly/odi-stories-list

The slides will be made available online after the course.

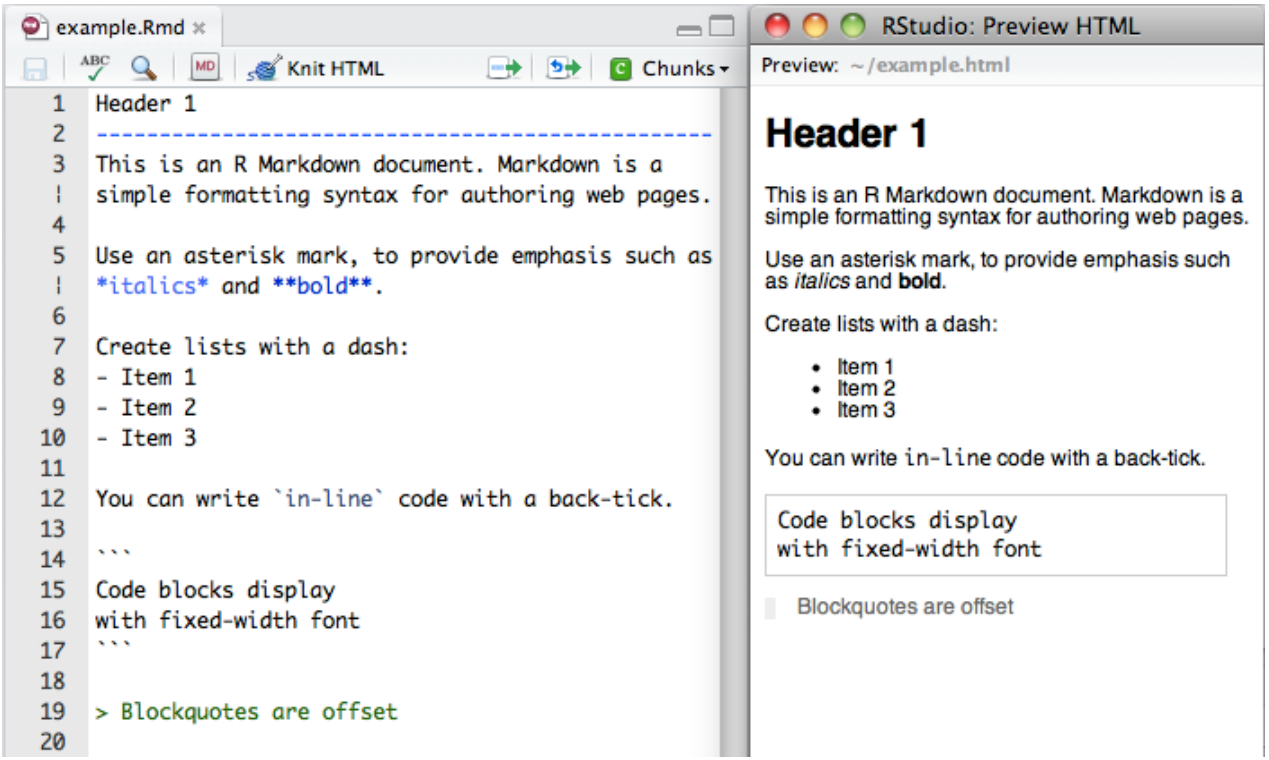
Appendix

http://www.cookbook-r.com





R Markdown



http://rpubs.com/statshero