Visualization in **R**

Evan Fields

OR Software Tools 2016

IAP 2016

Outline

Today we'll see:

- Motivation (10 minutes)
- Introduction to `ggplot2` about 90 minutes)
- Introduction to `ggmap` (about 30 minutes)
- Exploratory data analysis (about 30 minutes)

Anscombe's quartet

x1	10	8	13	9	11	14	6	4	12	7
x2	10	8	13	9	11	14	6	4	12	7
x3	10	8	13	9	11	14	6	4	12	7
x4	8	8	8	8	8	8	8	19	8	8
y1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82
y2	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.1	9.13	7.26
y3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42
y4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.5	5.56	7.91

Anscombe's quartet

var	mean	variance
x1	9	11
x2	9	11
x3	9	11
x4	9	11
y1	≈ 7.5	≈ 4.12
y2	≈ 7.5	≈ 4.12
y3	≈ 7.5	≈ 4.12
y4	≈ 7.5	≈ 4.12

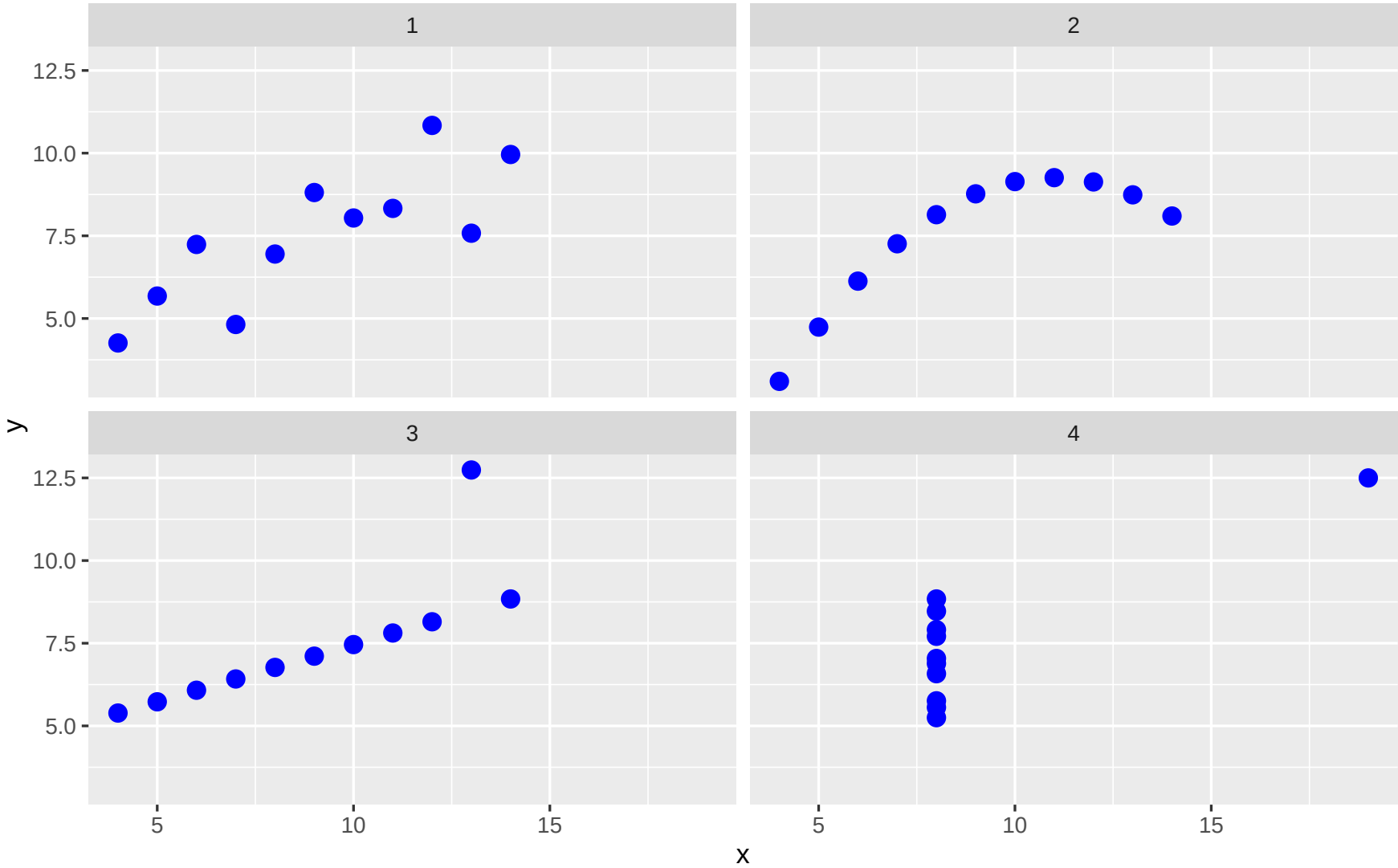
Anscombe's quartet

Almost identical linear regressions, to two decimal places:

$$\widehat{y1} = 3 + .5x1$$
$$\widehat{y2} = 3 + .5x2$$
$$\widehat{y3} = 3 + .5x3$$
$$\widehat{y4} = 3 + .5x4$$

And all four regressions have a r^2 of .67, also to two decimal places!

Anscombe's quartet



ggplot2

"A grammar of graphics is a tool that enables us to concisely describe the components of a graphic. Such a grammar allows us to move beyond named graphics (e.g., the "scatterplot") and gain insight into the deep structure that underlies statistical graphics."
– Hadley Wickham

ggplot2

A visualization is a mapping from data to visual properties. In ggplot2 we specify these explicitly.

ggplot2 plots have 3 basic components:

- The data that the plot will use
- The mapping from data to visual properties
- The geometry that will be drawn with these visual properties

Exercise 1

Make a scatterplot of fare vs. tip. Make sure your plot is nicely labeled and play with point color, shape, etc.

Exercise 2

Use chaining to make a line plot of passenger count vs. average fare. This will require some commands learned last Thursday such as `group_by` and `summarize`.

Bonus: Filter out any trips with 0 passengers or 0 tip.

Exercise 3

Create a column `mult_pass` which is true/false for whether passenger_count > 1 and make side-by-side and overlaid histograms exploring fare paid vs `mult_pass`.

Ensure sure the bin width is something reasonable. You might try stacking the histograms vertically by using `mult_pass ~ .` instead of `. ~ mult_pass`

Bonus: do some filtering to get more meaningful results.

Exercise 4

Plot the dropoff locations on the map of NYC. Use some transparency to improve readability in Manhattan. Color the points by trip distance.

Bonus: use `get_map` to get a map zoomed in on a particular area of NYC.

Exercise 5

Plot the `loess` trend. Change the size, color, etc. for enhanced clarity.

Bonus: fix the axis scales.

Make several plots with different random samples from the whole trips table. Do you notice anything interesting for trips of about 1 mile in length?

Exercise 6

Explore visually how travel distance changes with time of day.