Lecture 10:
Scatter Plots

April 6, 2020

Overview

Course Administration

1. Looking forward
   - Lecture 11: consultations. Sign up!
   - Lecture 12: storytelling and RShiny
   - Lecture 13: video presentations
   - Lecture 14: consultations. Sign up!
2. More on video presentations by email
3. Office hours have changed. Now
   - Tuesdays, 1:15 to 4:45
   - Thursdays, 8:15 to 9:15
   - Meet me in my WebEx room: see email for location
4. Will try to have all remaining assignment grades for you to check by next Monday on Jill's sheet
5. Paper due Monday May 4 by 5 pm to google drive. **Do not be late.**
6. Anything else?

The Next (and Last) Week of Good, Bad, and Ugly

Finder	Commenter
Danielle C.	Erik C.
Aaron K.	Caitlyn V.
Caitlyn V.	Lauren G.

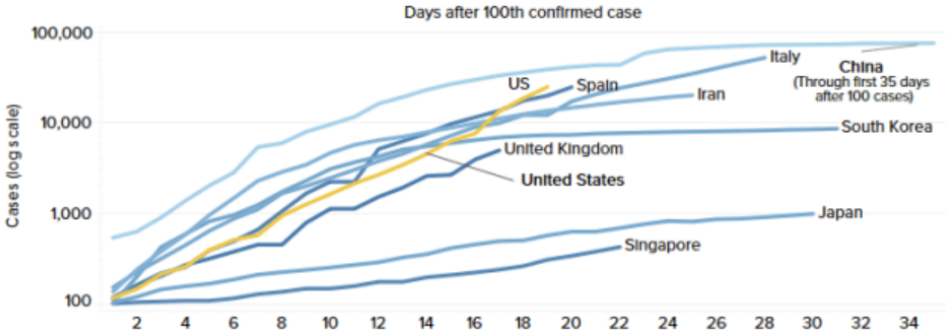
This Week's Good Bad and Ugly

Finder	Commenter
Lydia G.	Aaron K.
David N.	Basia D.

Lydia's Example. Aaron Discusses.

Total confirmed cases

The number of new cases of coronavirus began to slow in China and South Korea about three weeks after the first 100 cases had been reported. In the U.S., the number of reported cases has been slowed by a shortage of diagnostic test kits. That could bring a surge of new cases as test kits become available.



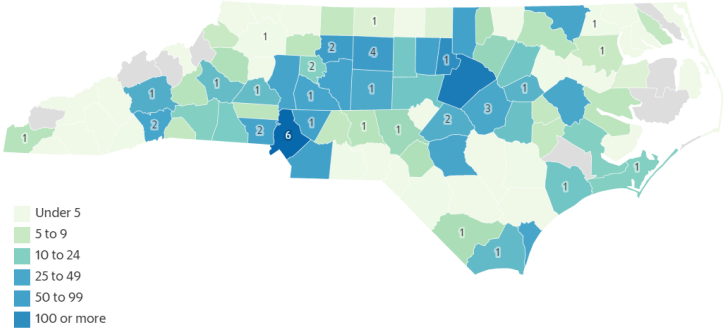
SOURCE: CNBC analysis of data from Johns Hopkins University Center for Systems Science and Engineering
(Selected countries 10 days after 100th case, as of March 21)

CNBC

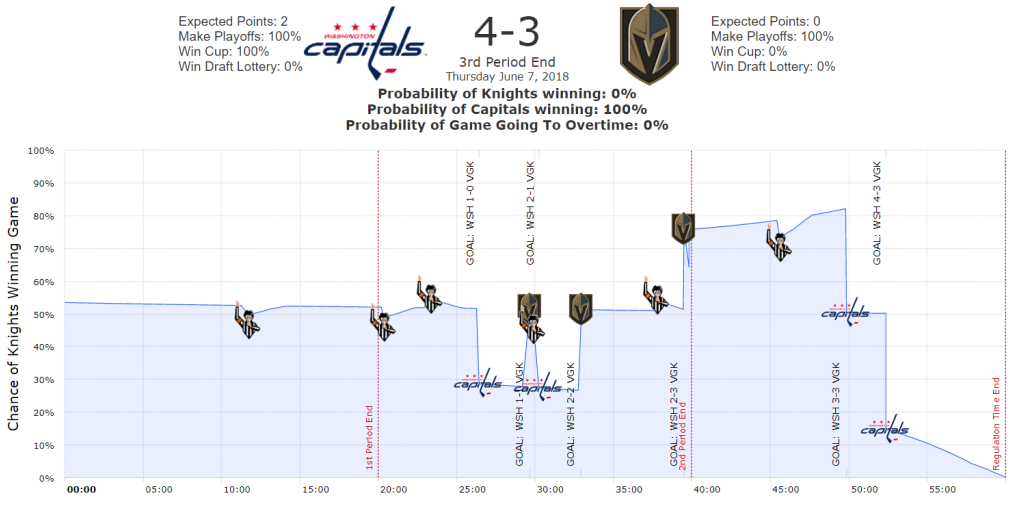
Kaila's Example. Dallas Discusses.

NC CORONAVIRUS CASES

Number of reported coronavirus cases by county as reported by NC DHHS and county health departments. Figures for the number of people who have recovered after testing positive are not available. Not all cases of COVID-19 are tested, so this does not represent the total number of people who have or had the virus.The number in the county represents the number of reported deaths due to the virus.



David's Example. Basia Discusses.



From *Monkey Puck*

Origins of Scatter Plots

What is a Scatterplot?

What is a Scatterplot?

- Plots values of two different variables on the same chart

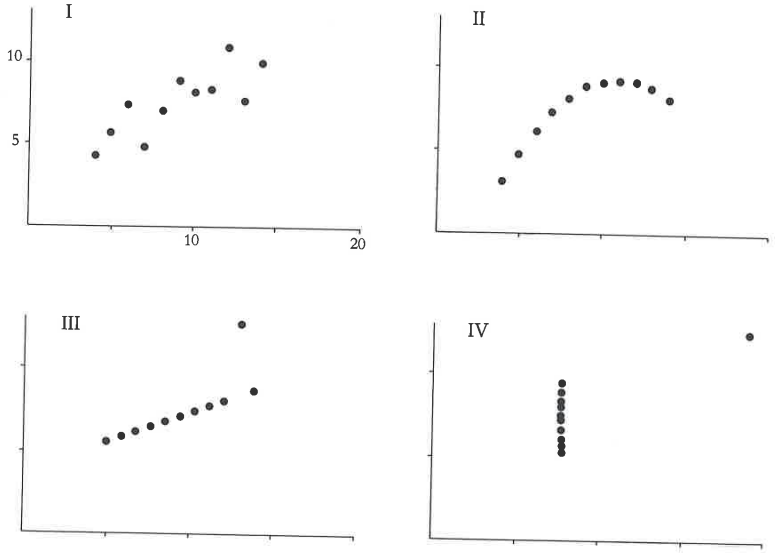
What is a Scatterplot?

- Plots values of two different variables on the same chart
- Shows correlation between two items

A Reminder and Example: Anscombe's Quartet

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

A Reminder and Example: Anscombe's Quartet



What Makes a Scatter Plot Different From All Other Plots?

(That We have Studied) – from Friendly and Denis, 2005

What Makes a Scatter Plot Different From All Other Plots?

(That We have Studied) – from Friendly and Denis, 2005

- it is fundamentally 2-D

What Makes a Scatter Plot Different From All Other Plots?

(That We have Studied) – from Friendly and Denis, 2005

- it is fundamentally 2-D
- a line graph is sort of 2-D, but only really for time

What Makes a Scatter Plot Different From All Other Plots?

(That We have Studied) – from Friendly and Denis, 2005

- it is fundamentally 2-D
- a line graph is sort of 2-D, but only really for time
- everything else we've studied is either a categorical relationship
  - bar chart

What Makes a Scatter Plot Different From All Other Plots?

(That We have Studied) – from Friendly and Denis, 2005

- it is fundamentally 2-D
- a line graph is sort of 2-D, but only really for time
- everything else we've studied is either a categorical relationship
  - bar chart
- or 1-D
  - histogram

What Makes a Scatter Plot Different From All Other Plots?

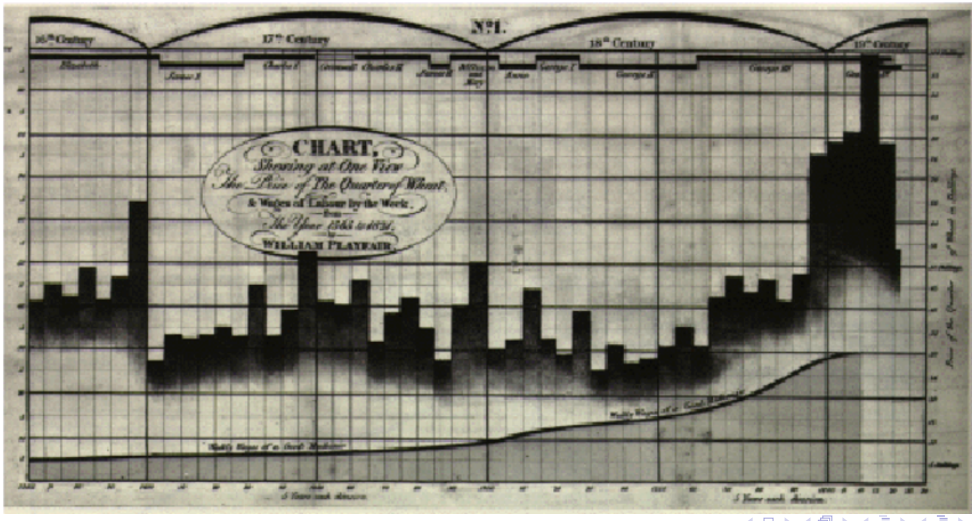
(That We have Studied) – from Friendly and Denis, 2005

- it is fundamentally 2-D
- a line graph is sort of 2-D, but only really for time
- everything else we've studied is either a categorical relationship
  - bar chart
- or 1-D
  - histogram

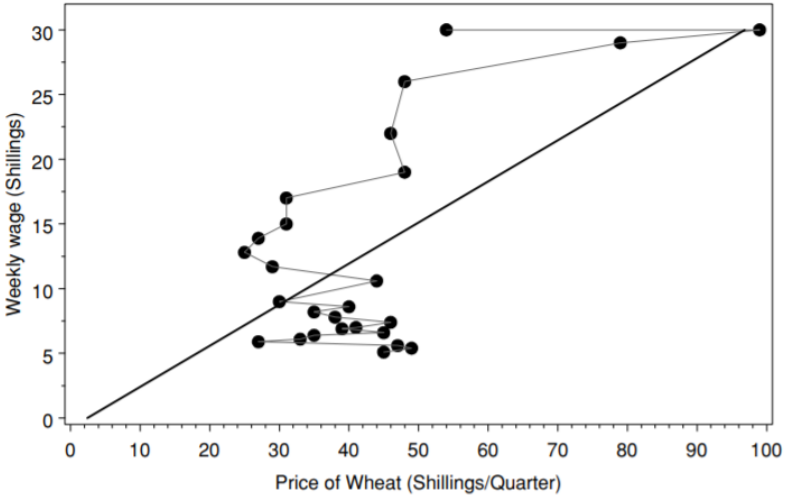
Map is the closest analogue to a scatter: points in (x, y) space

Scatters Are the Most Modern of Graphs We Study

Playfair approached, but didn't get to this form. Wages as line; wheat prices as bars.

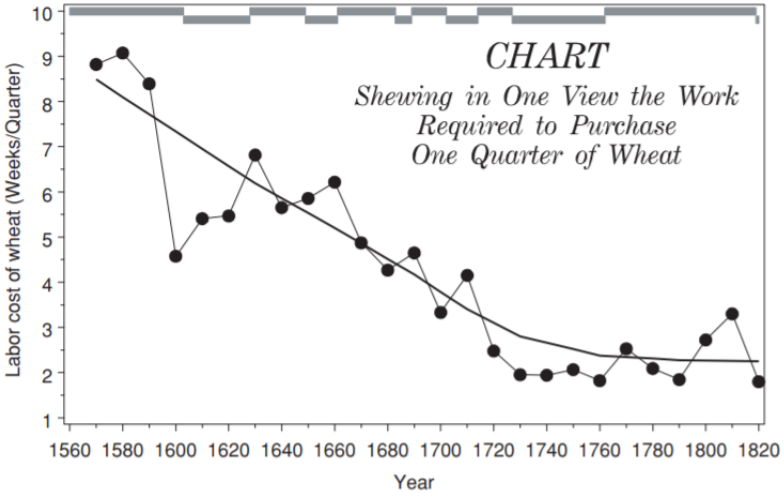


Playfair's Graph as a Proper Scatter



Connecting line is time

Revision of Playfair Makes the Key Point – But is Not a Scatter



One of the First Scatterplots: 1886

The Graph

- aims to predict one variable from the other
- has no time dimension
- notes density of observations

One of the First Scatterplots: 1886

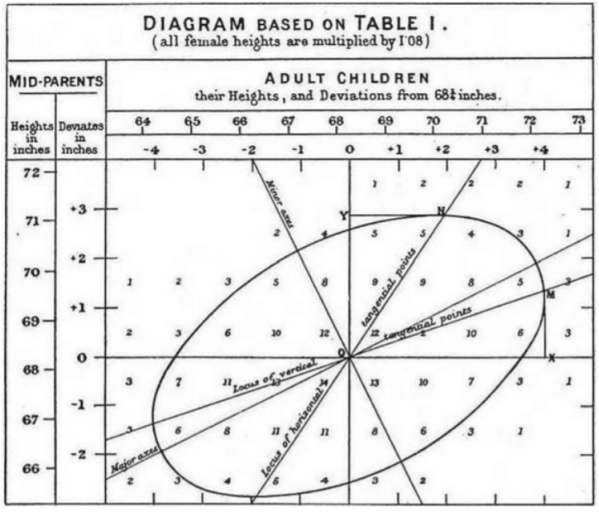
The Graph

- aims to predict one variable from the other
- has no time dimension
- notes density of observations

The Author: Francis Galton

- a measurer of all things: weather, height, etc
- invented or first described
  - the questionnaire
  - standard deviation
  - regression to the mean
- and the developer of eugenics

Galton's Scatter



Galton, 1886.

How and When to Use Scatters

Pros and Cons of Scatters

Most common type of graph for academic presentation

Pros and Cons of Scatters

Most common type of graph for academic presentation

Pros

- Can clearly and compellingly show a bivariate relationship
- Shows relationship throughout the distribution

Pros and Cons of Scatters

Most common type of graph for academic presentation

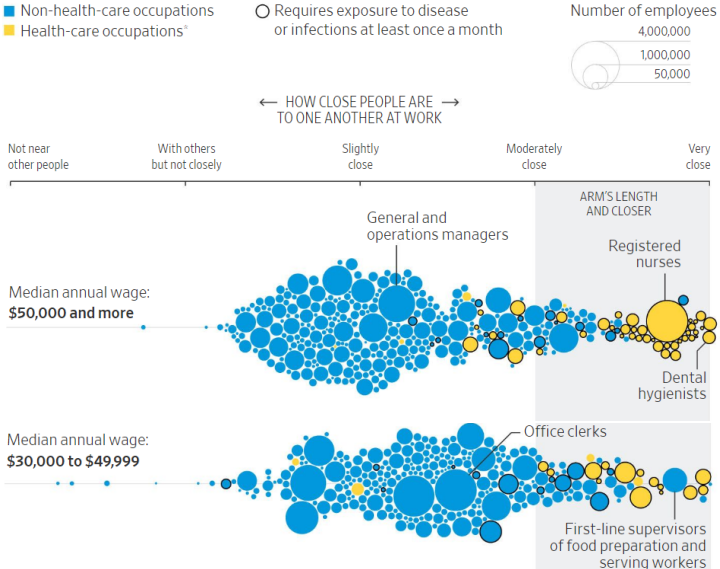
Pros

- Can clearly and compellingly show a bivariate relationship
- Shows relationship throughout the distribution

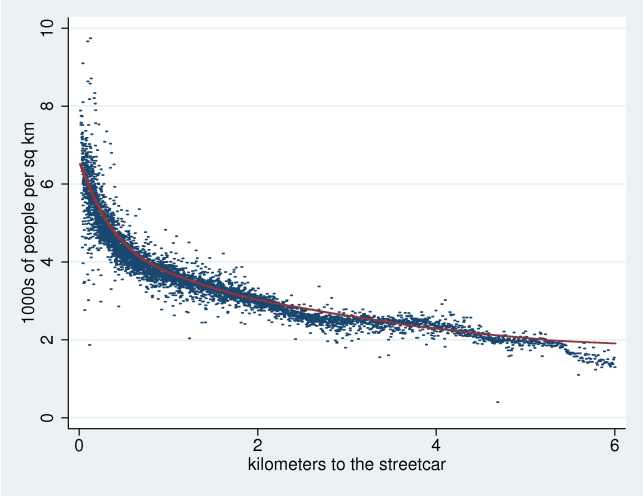
Cons

- Requires the audience to think about the relationship
- Sometimes too complicated for policy communication
- Can obscure relationships that do exist

This Should be a Scatter But Was Not



My Best Ever Scatter



How Can You Annotate a Scatter?

How Can You Annotate a Scatter?

- best fit lines
- ovals
- colors
- call out individual items

How to Deal with Issues of Multiple Variables

1. If they are in the same units?

How to Deal with Issues of Multiple Variables

1. If they are in the same units? graph on the same scale
2. If they are in different units?

How to Deal with Issues of Multiple Variables

1. If they are in the same units? graph on the same scale
2. If they are in different units?
   - can use two axes, but rarely a good idea – why?
   - plot on two charts side-by-side
   - do you want side-by-side vertical or horizontal?
3. If you have many different variables to show?

Small Multiples

When do you use them?

- Multiple variables to show
- Too much for one graph
- In presentations, usually helpful to explain one part first

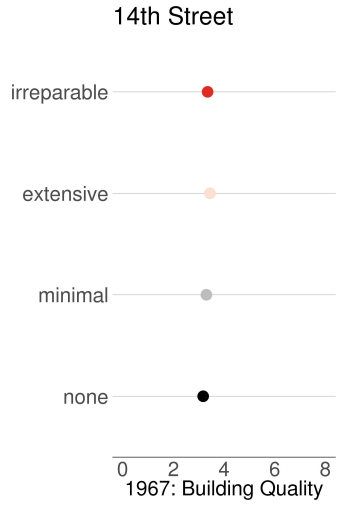
There is an implicit assumption that all graphs use the same scale.

How Beyonce Exploits the Power of Small Multiples

With thanks to Vibe.

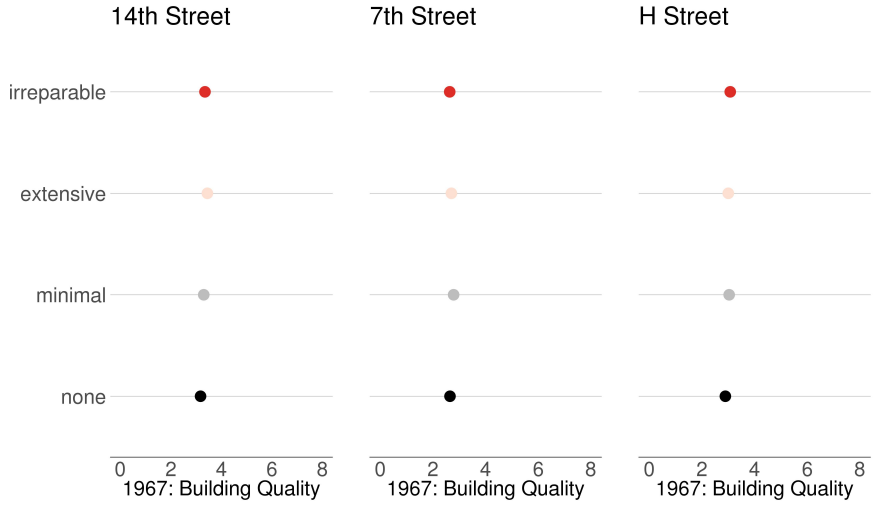
My Small Multiples

Destruction Roughly Even by 1967 Quality



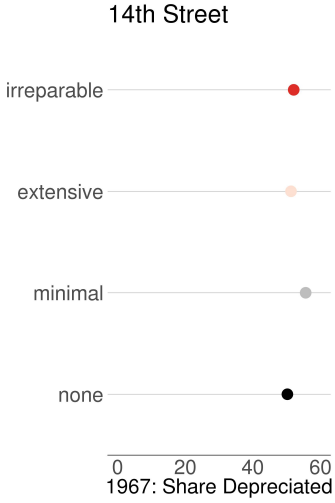
My Small Multiples

Destruction Roughly Even by 1967 Quality



My Small Multiples

Destruction Roughly Even by 1967 Depreciation



My Small Multiples

Destruction Roughly Even by 1967 Depreciation

R Notes

Today in R: Line Charts and De-Bugging

1. Scatter plots: `geom_point()`
2. Segments: `geom_segment()`
3. Small multiples
4. Instead of a loop: Use vector power

1. Scatter plots

```
p1 <- ggplot() +
  geom_point(data = df,
             mapping = aes(x = xvar, y = yvar))
```

Scatter plots: Shapes

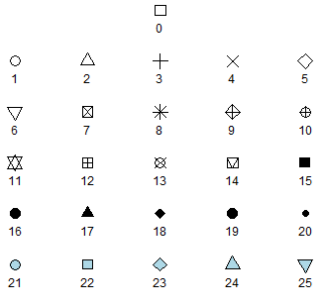


Figure 1:

Scatter plots: Shapes

Figure 1:

```
p1 <- ggplot() +
  geom_point(data = df,
             mapping = aes(x = xvar, y = yvar),
```

Scatter plots: One color

```
p1 <- ggplot() +
  geom_line(data = polys,
            mapping = aes(x = xvar, y = yvar),
            color = "COLOR.NAME")
```

Scatter plots: Colors by Group

```
p1 <- ggplot() +
  geom_line(data = polys,
            mapping = aes(x = xvar, y = yvar),
            color = VARIABLE)
```

Scatter plots: Colors by Group

```
p1 <- ggplot() +
  geom_line(data = polys,
            mapping = aes(x = xvar, y = yvar),
            color = VARIABLE)
```

- To show colors by a variable
- You can specify colors in

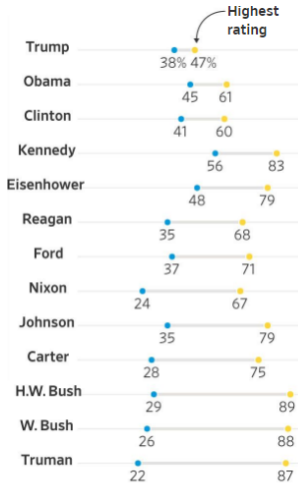
```
  scale_colour_manual(values=c('A'='grey',
                               'E'='red',
                               'F'='blue'))
```

Scatter plots: Calling out Regions

- best fit line: use cautiously
  `geom_smooth(method = lm, se = FALSE)`
- best fit curve: same
  `geom_smooth(se = FALSE)`
- best fit curve: with shaded error region
  `geom_smooth()`
- annotations
  `geom_rect()` `geom_segment()`

2. Drawing Segments

This is a scatterplot with segments!



Code Segments

```
s2 <- ggplot() +
      geom_segment(data = df,
                   mapping = aes(x = VARIABLE1,
                                 xend = VARIABLE2,
                                 y = VARIABLE3,
                                 yend = VARIABLE4))
```

Code Segments

```
s2 <- ggplot() +
      geom_segment(data = df,
                   mapping = aes(x = VARIABLE1,
                                 xend = VARIABLE2,
                                 y = VARIABLE3,
                                 yend = VARIABLE4))
```

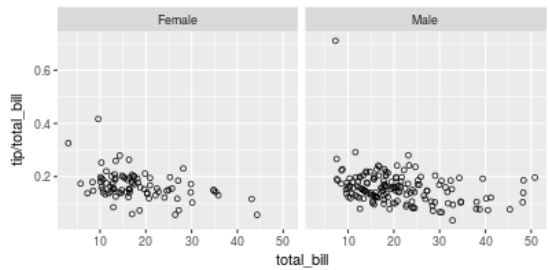
There is also geom_curve for brave people

3. Small Multiples, or Facets

```
facet_grid(rows = vars(VARIABLE))
```

3. Small Multiples, or Facets

```
facet_grid(rows = vars(VARIABLE))
```



Thanks to Winston Chang.

Facet Columns

```
facet_grid(cols = vars(VARAIBLE))
```

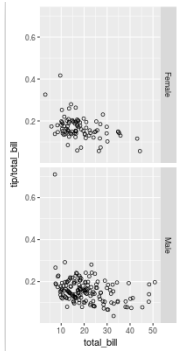


Figure 3:

4. Avoiding a Loop

Suppose you want to do this many times

```
df$ln.x <- log(df$x)
```

4. Avoiding a Loop

Suppose you want to do this many times

```
df$ln.x <- log(df$x)
```

This does not work!

```
tolog <- c(x,y,z)
for(i in tolog){
  df$ln.i <- log(df$i)
}
```

and you can't fix it up with `eval(parse())` either.

The Elegant Solution

```
tolog <- c("x","y","z")
df[paste0("ln.",tolog)] <- log(df[tolog])
```

The Elegant Solution

```
tolog <- c("x","y","z")
df[paste0("ln.",tolog)] <- log(df[tolog])
```

Recall:

$$y = log_b(x)$$

and

$$x = b^y$$

The Elegant Solution in Action

```
df <- data.frame(x = c(1, 2, 3),
                 y = c(10, 20, 30),
                 z = c(100, 200, 300))
```

The Elegant Solution in Action

```
df <- data.frame(x = c(1, 2, 3),
                 y = c(10, 20, 30),
                 z = c(100, 200, 300))
```

```
df
```

```
##   x  y   z
## 1 1 10 100
## 2 2 20 200
## 3 3 30 300
```

The Elegant Solution in Action

```
df <- data.frame(x = c(1, 2, 3),
                 y = c(10, 20, 30),
                 z = c(100, 200, 300))
tolog <- c("x","y","z")
df[paste0("ln.",tolog)] <- log(df[tolog])
df
```

```
##   x  y   z      ln.x     ln.y     ln.z
## 1 1 10 100 0.0000000 2.302585 4.605170
## 2 2 20 200 0.6931472 2.995732 5.298317
## 3 3 30 300 1.0986123 3.401197 5.703782
```

Next Lectures

- Consultations
- Video presentations due April 27