Lecture 8:
Scatter Plots and Color

July 6, 2022

Overview

Course Administration

Good, Bad and Ugly

Scatter Plot Origins

How and When to Use Scatters

Showing Multiple Variables or Variations

Color

R Notes

Course Administration

1. Looking forward
   - Lecture 9, July 13: Storytelling and interactive plots
   - Lecture 10, July 20: 5 minute presentations
   - July 27: policy brief due
2. Anything else?

Next Week's Assignment

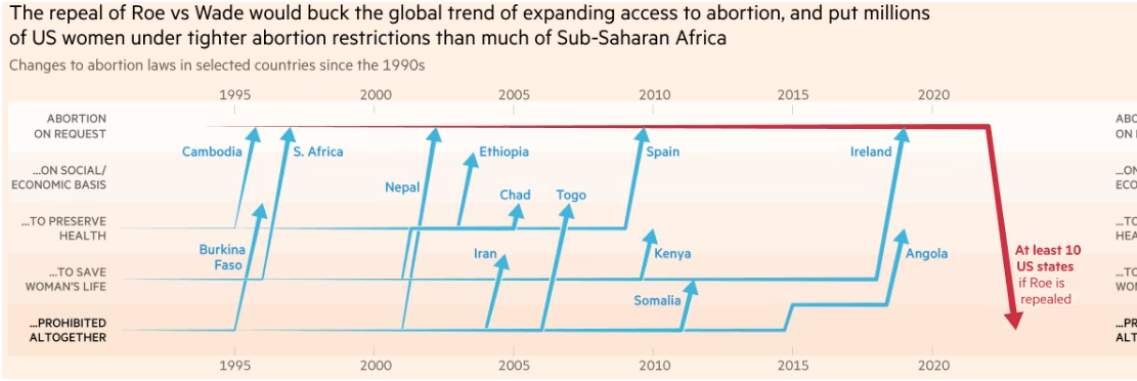
Find a scatterplot. Post link to google sheet by Wednesday noon.

Finder	Commenter
Jarred	Linsi
Sarah	Brandon
Esther	Dayo

This Week's Good Bad and Ugly

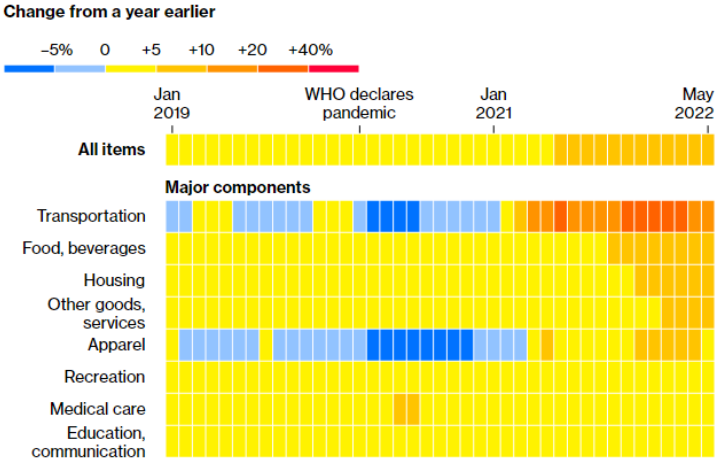
Finder	Commenter
Linsi G.	Richa
Brandon	Esnold
Dayo	Esther

Richa on Linsi's Graphic



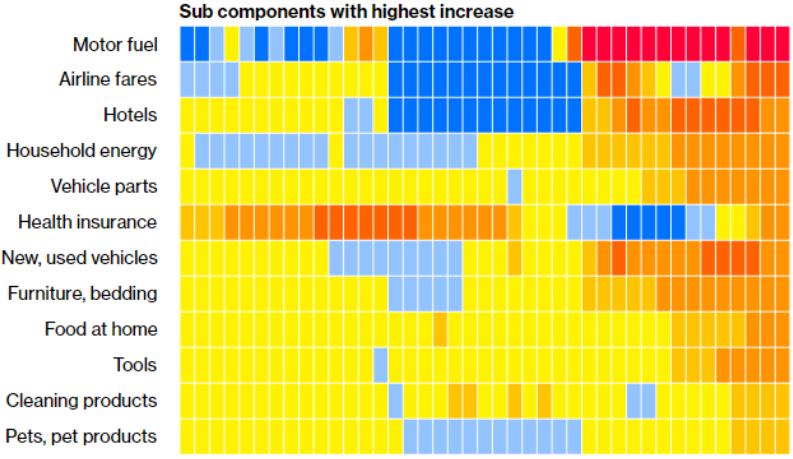
Burn-Murdoch, John, "Repeal of Roe risks exacerbating the US's most shameful statistic," *Financial Times*, May 5, 2022. [link]

Esnold on Brandon's Chart, 1 of 2



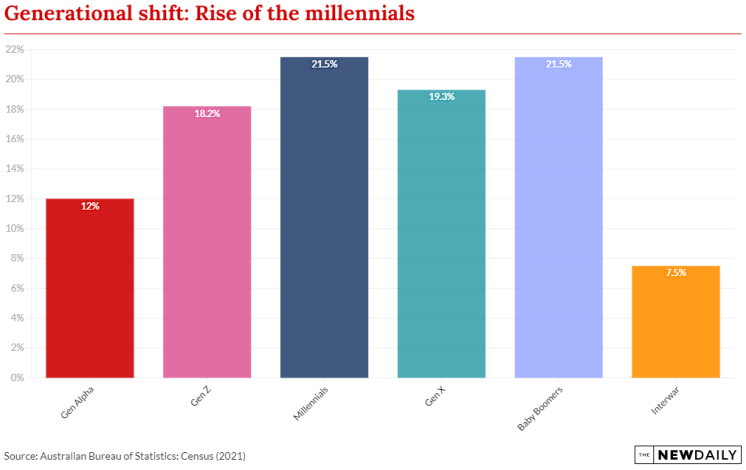
Burgess, Robert et al, "How close are we really to 1970s style inflation?," *Bloomberg Opinion*, June 11, 2022. [link]

Esnold on Brandon's Chart, 2 of 2



Burgess, Robert et al, "How close are we really to 1970s style inflation?," *Bloomberg Opinion*, June 11, 2022. [link]

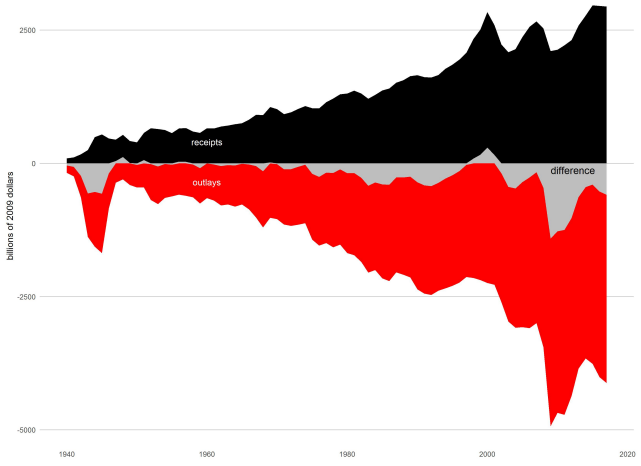
Esther on Dayo's Graphic



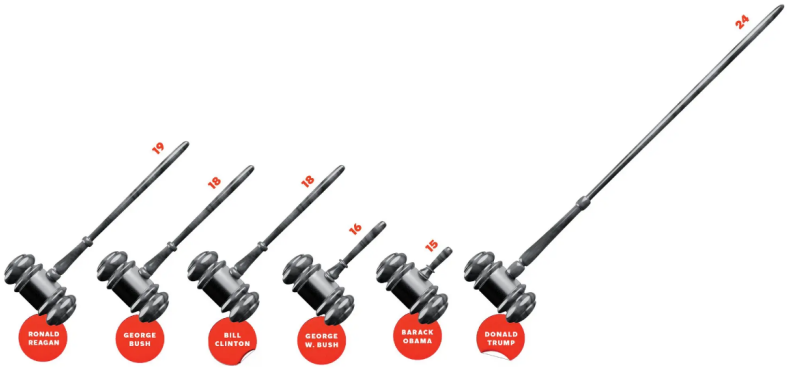
Elmas, Matthew and Zara Falkiner-Rose, "These five Census graphs tell a story about a rapidly changing Australia," *The New Daily*, June 28, 2022. [link]

My Surplus Chart

My Surplus Chart



Esnold's Terrible *NYT* Graphic



Zengerle, Jason. "How the Trump Administration is Remaking the Courts," *NYT* August 22, 2018. [link].

In Table Form

	year	judges	gavel end	gavel top
	(1)	(2)	(3)	(4)
Reagan	1980	19	3.5	3
Bush 1	1988	18	2.875	2.375
Clinton	1992	18	2.875	2.375
Bush 2	2000	16	1.75	1.25
Obama	2008	15	1.125	0.625
Trump	2016	24	6.375	5.875

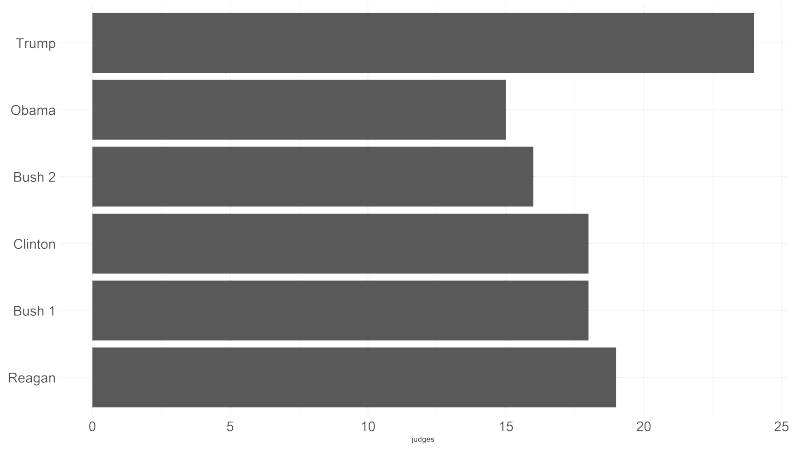
In Table Form

	year	judges	gavel end	gavel top	(3) / (2)
	(1)	(2)	(3)	(4)	(5)
Reagan	1980	19	3.5	3	0.18
Bush 1	1988	18	2.875	2.375	0.16
Clinton	1992	18	2.875	2.375	0.16
Bush 2	2000	16	1.75	1.25	0.11
Obama	2008	15	1.125	0.625	0.08
Trump	2016	24	6.375	5.875	0.27

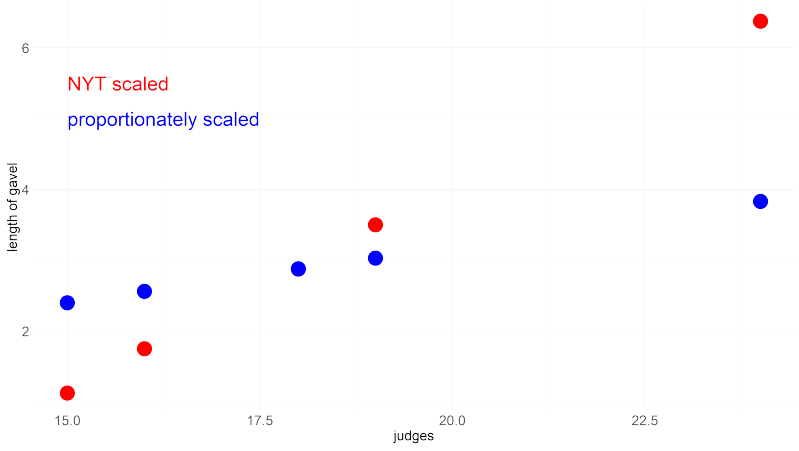
In Table Form

	year	judges	gavel end	gavel top	(3) / (2)	(5)*(2) if 18
	(1)	(2)	(3)	(4)	(5)	(6)
Reagan	1980	19	3.5	3	0.18	3.03
Bush 1	1988	18	2.875	2.375	0.16	2.88
Clinton	1992	18	2.875	2.375	0.16	2.88
Bush 2	2000	16	1.75	1.25	0.11	2.56
Obama	2008	15	1.125	0.625	0.08	2.4
Trump	2016	24	6.375	5.875	0.27	3.83

As as a Graphic, 1 of 2



As as a Graphic, 1 of 2



This Lecture

1. Scatter plot definition and origins
2. How and when to use scatters
3. Small multiples
4. Color
5. R stuff

Scatter Plot: Definition and Origins

What is a Scatter Plot?

What is a Scatter Plot?

- Plots values of two different variables on the same chart

What is a Scatter Plot?

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

What is a Scatter Plot?

- Plots values of two different variables on the same chart
- Shows correlation between two variables
- Can also show distribution of each variable

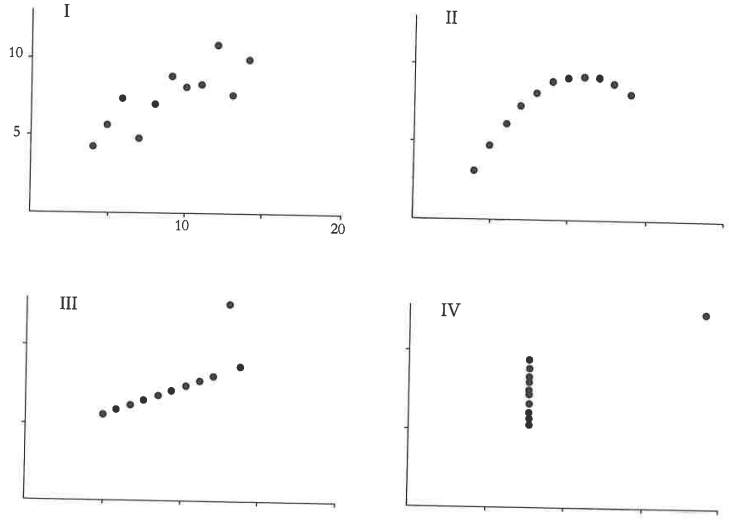
A Reminder and Example: Anscombe's Quartet

Same mean, same variance

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

Same mean, same variance



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

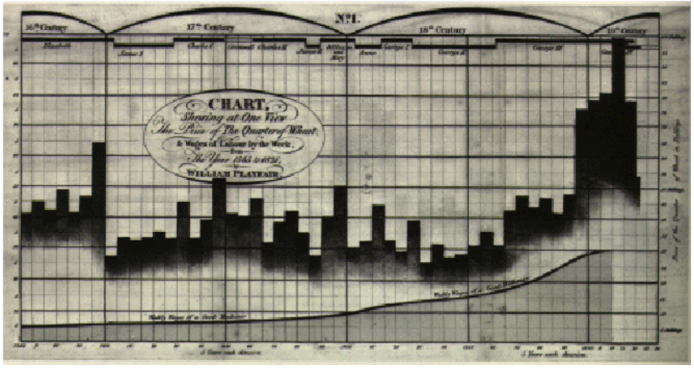
What Makes a Scatter Plot Different From All Other Plots?

(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

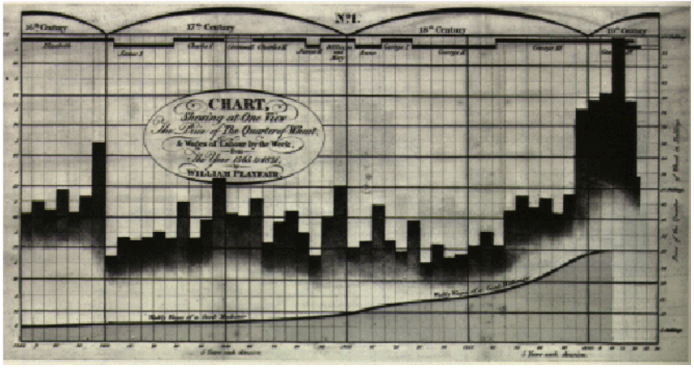
Map is the closest analogue to a scatter: points in (x, y) space

Scatters Are the Most Modern of Graphs We Study



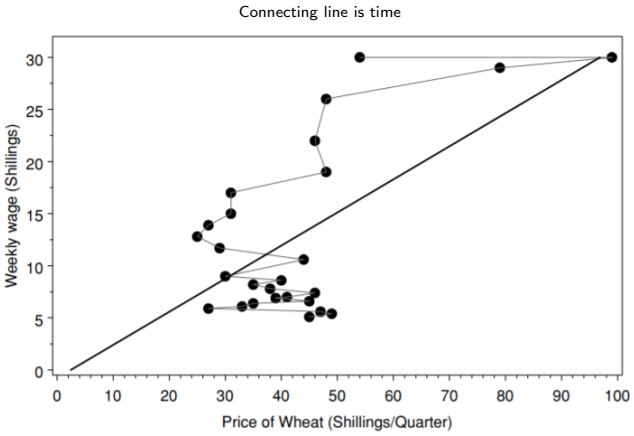
- What is this graph?
  - two y axes
  - wages in line
  - price of wheat in bars
  - horizontal axis is time
- What is the goal of this graph?

Scatters Are the Most Modern of Graphs We Study



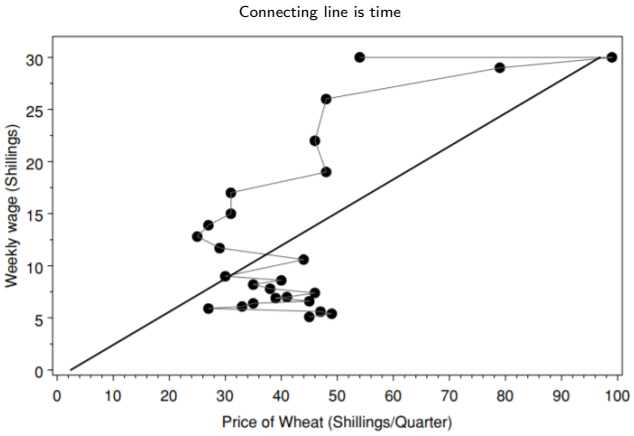
- What is this graph?
  - two y axes
  - wages in line
  - price of wheat in bars
  - horizontal axis is time
- What is the goal of this graph?
  - show that purchasing power increases over time
  - is it clear?

Playfair's Graph as a Proper Scatter



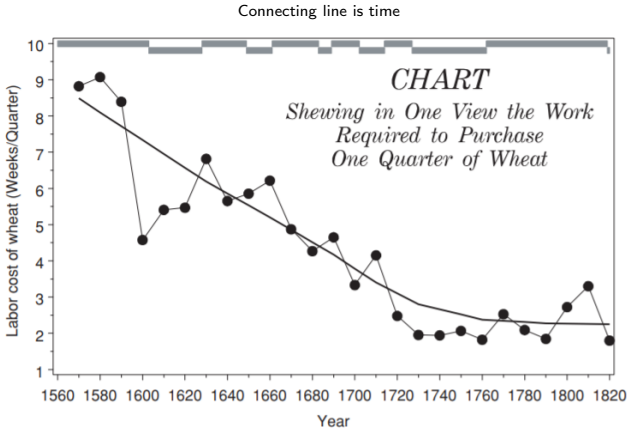
- What is this graph?
  - price of wheat on x
  - wage on y
  - line connects by time

Playfair's Graph as a Proper Scatter



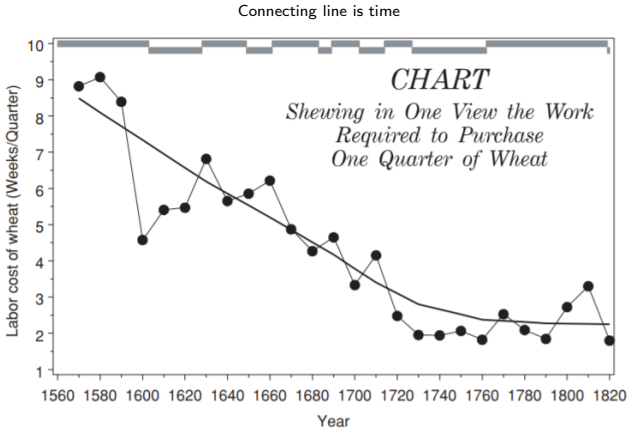
- What is this graph?
  - price of wheat on x
  - wage on y
  - line connects by time
- Why is this graph not too helpful?
  - you don't know when is when
  - no temporal point

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



- What is this graph?
  - time on x
  - on y, number of weeks required to purchase one quarter of wheat
  - line connects by time

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



- What is this graph?
  - time on x
  - on y, number of weeks required to purchase one quarter of wheat
  - line connects by time
- Why is this better?
  - line connects time and you can see it
  - makes the ratio for you
  - the ratio is the point!

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

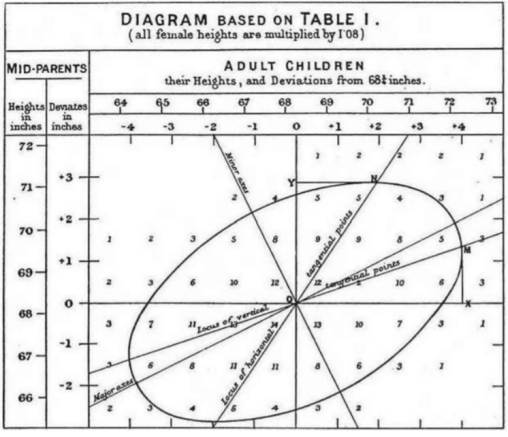
The Graph

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

The Author: Francis Galton

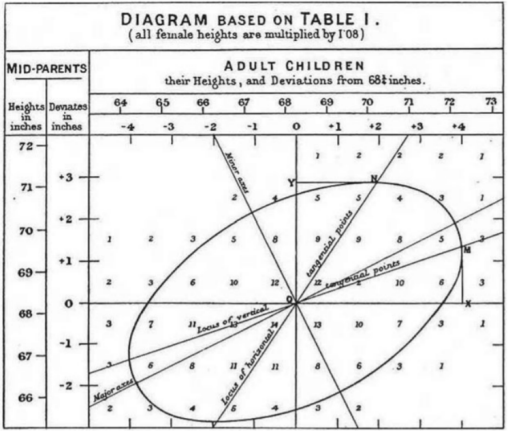
- a measurer of all things: weather, height, etc
- invented or first described
  - the questionnaire
  - standard deviation
  - regression to the mean
- and the developer of eugenics

Galton's Scatter



- What is this graph?
  - height of adult children on x
  - height of parents on y
  - numbers are the number of observations at each point

Galton's Scatter



- What is this graph?
  - height of adult children on x
  - height of parents on y
  - numbers are the number of observations at each point
- This is an early scatter
- Scatters are nor prevalent until the 1920s
- Still usually too complicated for most layperson communications

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

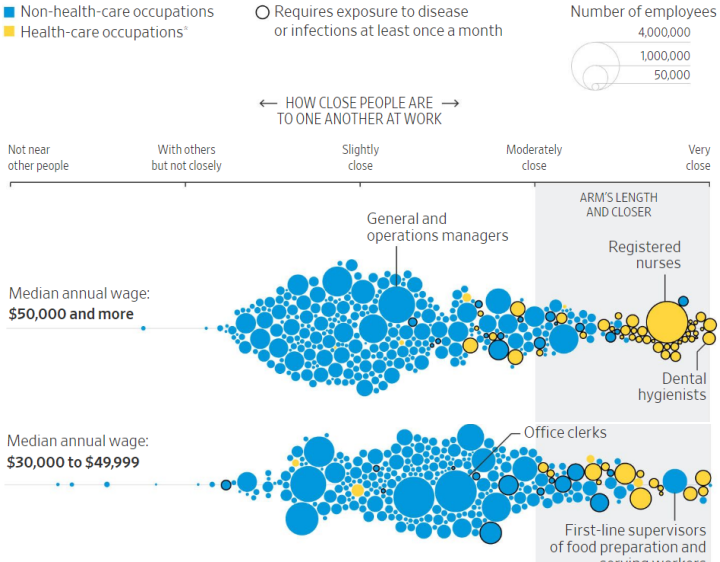
Pros

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

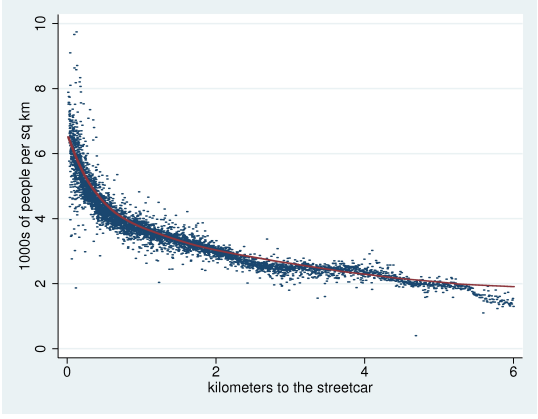
Cons

- 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



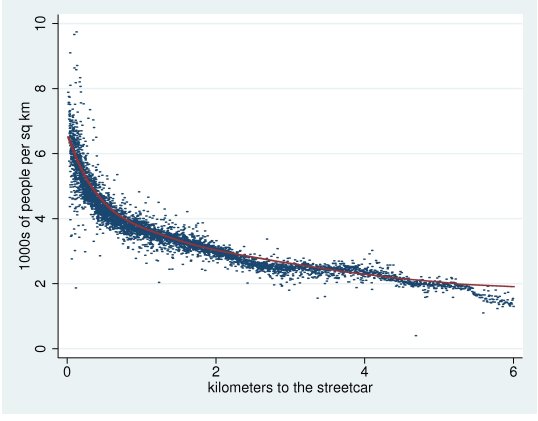
My Best Ever Scatter



What is it?

- Each point is
- average population density near about 400 land plots
- at a given distance from an old streetcar
- red line is a flexible regression line

My Best Ever Scatter

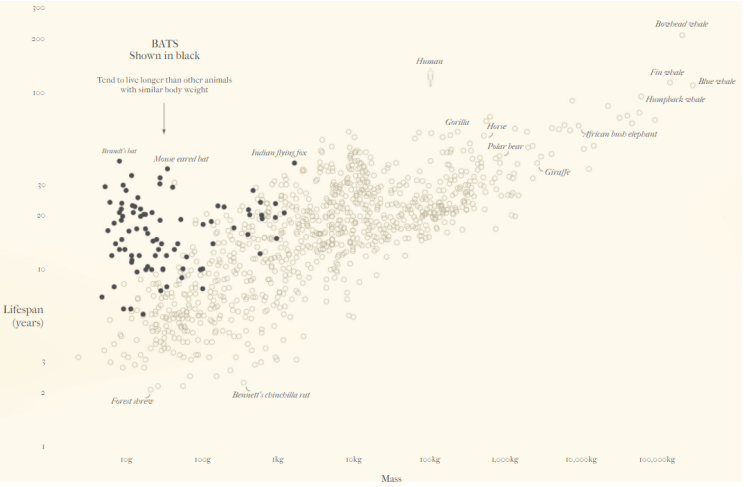


What is it?

- Each point is
- average population density near about 400 land plots
- at a given distance from an old streetcar
- red line is a flexible regression line

Data show the point

How Can You Annotate a Scatter?



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

Showing Multiple Variables or Variations

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

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?

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?

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?
   - see the next slide..

Small Multiples

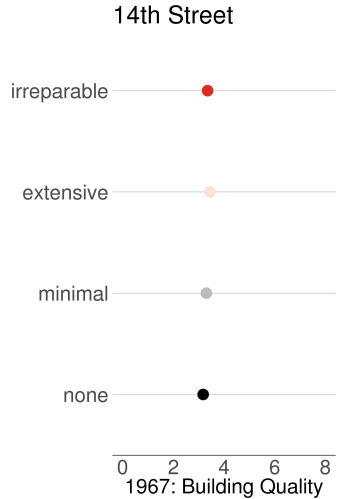
When do you use them?

- 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.

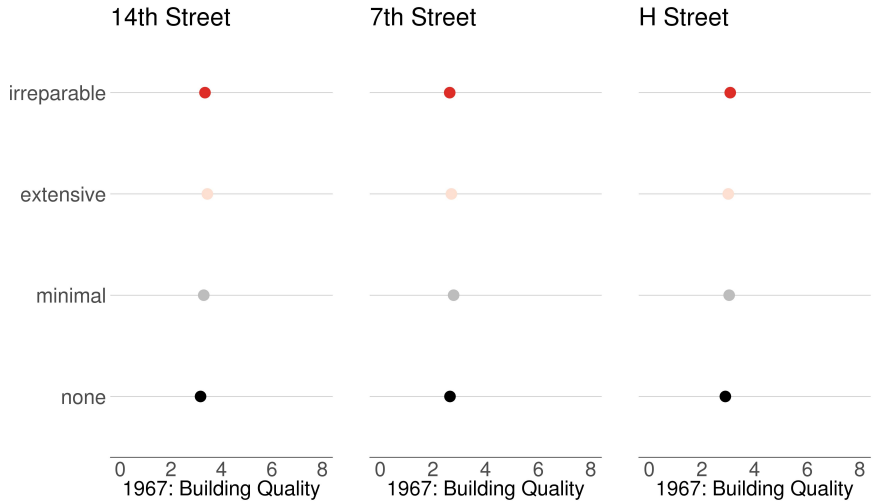
My Small Multiples

Destruction Roughly Even by 1967 Quality



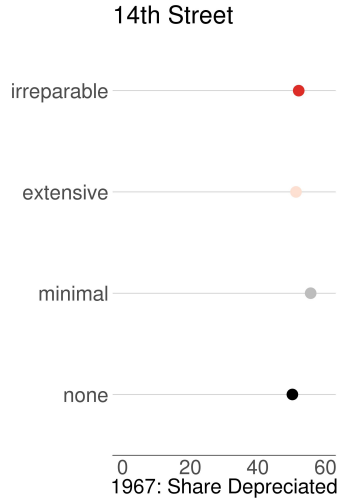
My Small Multiples

Destruction Roughly Even by 1967 Quality



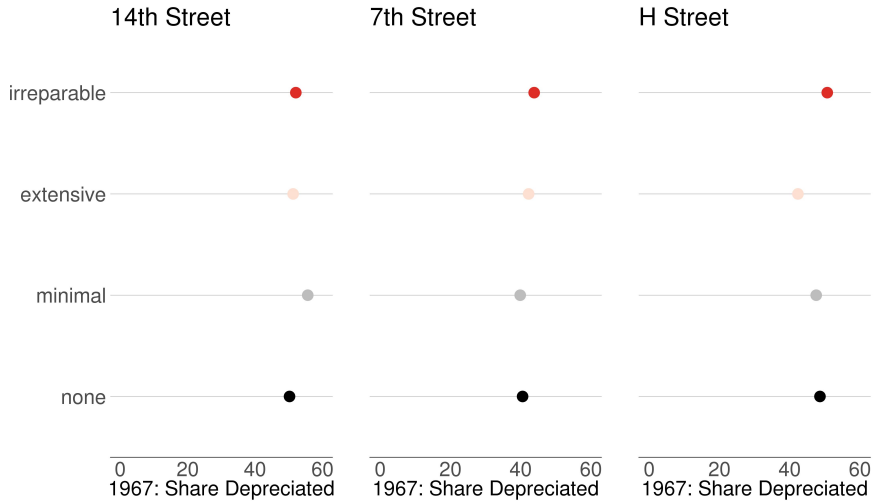
My Small Multiples

Destruction Roughly Even by 1967 Depreciation



My Small Multiples

Destruction Roughly Even by 1967 Depreciation



How Beyonce Exploits the Power of Small Multiples

With thanks to Vibe.

Color

Why Color?

Why Color?

- Preattentive processing

Why Color?

- Preattentive processing
- Allows you to avoid labels
  - put the name in the same color as the bar or line
- Allows you to make subtle connections across graphics

Why Color?

- Preattentive processing
- Allows you to avoid labels
  - put the name in the same color as the bar or line
- Allows you to make subtle connections across graphics

But don't use too many colors! We can't remember them.

Color Definitions

1. Hue
   - color or shade
   - based on perception
   - Ex: difference between ripe and not-so-ripe bananas

Color Definitions

1. Hue
   - color or shade
   - based on perception
   - Ex: difference between ripe and not-so-ripe bananas
2. Saturation
   - Mix of hue with white
   - Ex: colors of chocolate milk

Color Definitions

1. Hue
   - color or shade
   - based on perception
   - Ex: difference between ripe and not-so-ripe bananas
2. Saturation
   - Mix of hue with white
   - Ex: colors of chocolate milk
3. Brightness/Luminosity/Lightness
   - Mix of hue with darker shading
   - Ex: toast vs burned toast

Color Definitions

1. Hue
   - color or shade
   - based on perception
   - Ex: difference between ripe and not-so-ripe bananas
2. Saturation
   - Mix of hue with white
   - Ex: colors of chocolate milk
3. Brightness/Luminosity/Lightness
   - Mix of hue with darker shading
   - Ex: toast vs burned toast

Thank you to this helpful article for thoughts and pic; and to Amit Agarwal for examples.

Color on a Screen

- Computer mixes red, blue, green in a pixel
- Computer screens cannot generate all the colors you can see

Color on a Screen

- Computer mixes red, blue, green in a pixel
- Computer screens cannot generate all the colors you can see
- Colors look different on different screens

Color on a Screen

- Computer mixes red, blue, green in a pixel
- Computer screens cannot generate all the colors you can see
- Colors look different on different screens
- Represented multiple ways
  - RGB: (xxx,xxx,xxx). Aqua: (0, 255, 255)
  - Hex: #XXXXXX. Aqua: #00FFFF

Thank you NASA! [link]

Types of Color Schemes

Types of Color Schemes

1. Qualitative/Categorical

2. Sequential

3. Divergent

Types of Color Schemes

Types of Color Schemes

1. Qualitative/Categorical

2. Sequential

3. Divergent

For what?

1. discrete things

Types of Color Schemes

Types of Color Schemes

1. Qualitative/Categorical

2. Sequential

3. Divergent

For what?

1. discrete things
   - maps with categories
   - lines by type

Types of Color Schemes

Types of Color Schemes

1. Qualitative/Categorical

2. Sequential

3. Divergent

For what?

1. discrete things
   - maps with categories
   - lines by type
2. continuous values

Types of Color Schemes

Types of Color Schemes

1. Qualitative/Categorical

2. Sequential

3. Divergent

For what?

1. discrete things
   - maps with categories
   - lines by type
2. continuous values
   - dollar amounts
   - shares of population

Types of Color Schemes

Types of Color Schemes

1. Qualitative/Categorical

2. Sequential

3. Divergent

For what?

1. discrete things
   - maps with categories
   - lines by type
2. continuous values
   - dollar amounts
   - shares of population
3. continuous values where we care about breakpoint

Types of Color Schemes

Types of Color Schemes

1. Qualitative/Categorical

2. Sequential

3. Divergent

For what?

1. discrete things
   - maps with categories
   - lines by type
2. continuous values
   - dollar amounts
   - shares of population
3. continuous values where we care about breakpoint
   - up or down
   - high or low
   - hot or cold

How to Choose?

Recommendations

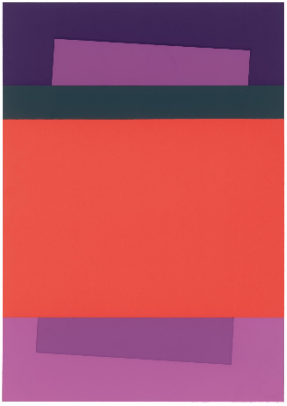
- ColorBrewer
- Data Color Picker
- Chroma.js Color Palette Helper
- Color Thief: makes palettes from images!
- Viz Palette: see your colors in fake graphics

Words of Warning: Color is Relative

- Culturally relative

Words of Warning: Color is Relative

- Culturally relative
- Optically relative



Interaction of Color, Josef Albers, Plate 4.4

R Notes

Today in R: Scatter Plots, Segments, Small Multiples and Vector Power

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

Scatter plots: Shapes


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

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_color_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()`

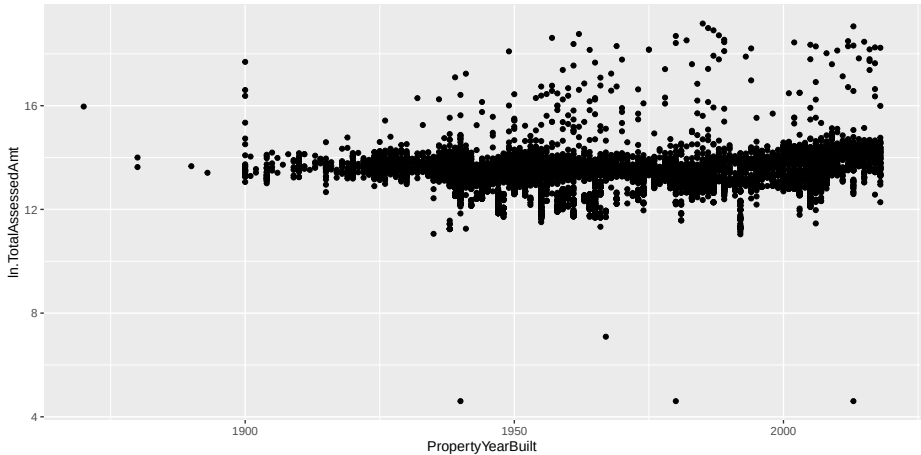
Some Examples With Property Data from Arlington, VA

- property data for Arlington County, VA
- observe attributes about properties
  - assessed value
  - year built
  - many other things

Some Examples With Property Data from Arlington, VA

```
p1 <- ggplot() +
  geom_point(data = arl.samp,
             mapping = aes(x = PropertyYearBuilt,
                           y = ln.TotalAssessedAmt))
```

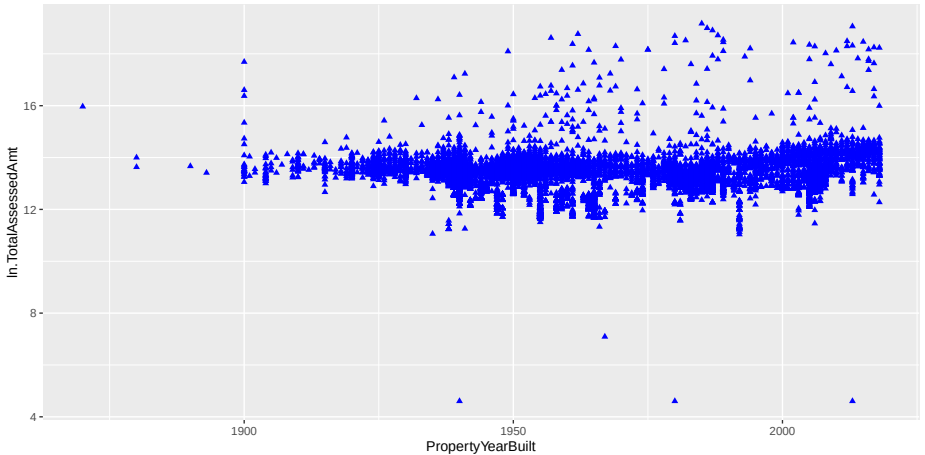
Some Examples With Property Data from Arlington, VA



Colors and Shape for Property Data from Arlington, VA

```
p2 <- ggplot() +
  geom_point(data = arl.samp,
             mapping = aes(x = PropertyYearBuilt,
                           y = ln.TotalAssessedAmt),
             color = "blue",
             shape = 17)
```

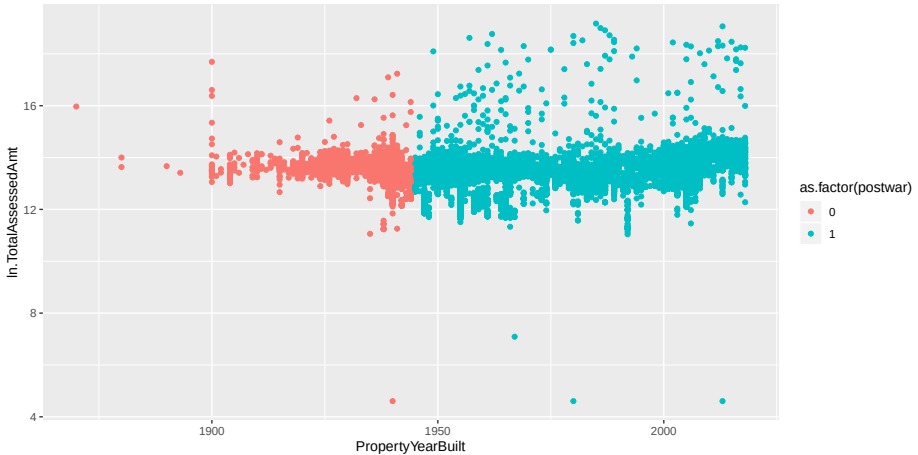
Colors and Shape for Property Data from Arlington, VA



Colors by Value for Property Data from Arlington, VA

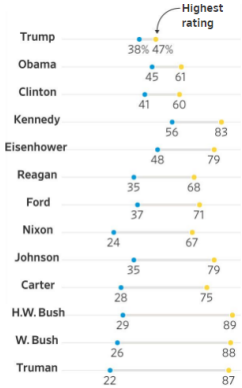
```
p2 <- ggplot() +
  geom_point(data = arl.samp,
             mapping = aes(x = PropertyYearBuilt,
                           y = ln.TotalAssessedAmt,
                           color = as.factor(postwar)),
             scale_color_manual = c("blue","red"))
```

Colors by Value for Property Data from Arlington, VA



2. Drawing Segments

This is a scatterplot with segments!



Thanks to WSJ.

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))
```

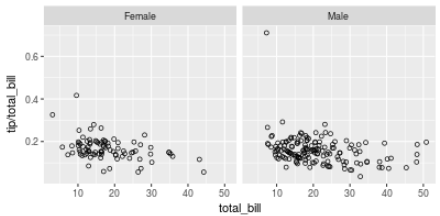
There is also `geom_curve` for brave people

3. Small Multiples, or Facets

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

3. Small Multiples, or Facets

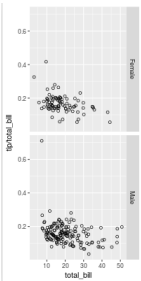
```
facet_grid(rows = vars(VARIABLE))
```



Thanks to Winston Chang.

Facet Columns

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



Or both.

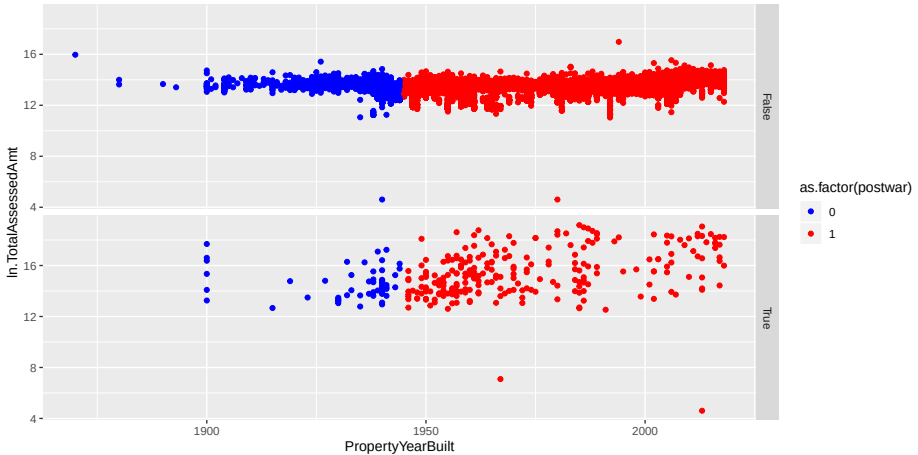
Faceting for Arlington

```
print(table(arl.samp$CommercialInd))
```

```
##
## False  True
## 11976   307
```

```
p2 <- ggplot() +
  geom_point(data = arl.samp,
             mapping = aes(x = PropertyYearBuilt,
                           y = ln.TotalAssessedAmt,
                           color = as.factor(postwar))) +
  scale_color_manual(values = c("blue","red")) +
  facet_grid(rows = arl.samp$CommercialInd)
```

Faceting for Arlington



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)
}
```

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])
```

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

- Presentations due online a few hours before you present
- More detailed instructions next week
- Final paper due July 27
- Office hours available – can schedule more as needed