Correlation

Communication Research Methods

Jennifer Pan

Assistant Professor
Department of Communication
Stanford University

February 17, 2016

Announcements

Announcements

- Monday Feb 29 Section and Tuesday Mar 1 Office Hours

Announcements

- Monday Feb 29 Section and Tuesday Mar 1 Office Hours
- Midcourse Survey

Announcements

- Monday Feb 29 Section and Tuesday Mar 1 Office Hours
- Midcourse Survey
- Grade Distribution

Midterm Feedback

Midterm Feedback

- 20 responses out of a class of 44 (selection bias?)

Midterm Feedback

- 20 responses out of a class of 44 (selection bias?)
- Lecture clear and engaging (8 people)

Midterm Feedback

- 20 responses out of a class of 44 (selection bias?)
- Lecture clear and engaging (8 people)
- Sections very helpful (4 people)

Midterm Feedback

- 20 responses out of a class of 44 (selection bias?)
- Lecture clear and engaging (8 people)
- Sections very helpful (4 people)
- Go slower when going through R in class (3 people)

Midterm Feedback

- 20 responses out of a class of 44 (selection bias?)
- Lecture clear and engaging (8 people)
- Sections very helpful (4 people)
- Go slower when going through R in class (3 people)
- Pset expectations not clear (5 people)

Midterm Feedback

- 20 responses out of a class of 44 (selection bias?)
- Lecture clear and engaging (8 people)
- Sections very helpful (4 people)
- Go slower when going through R in class (3 people)
- Pset expectations not clear (5 people)
  Median grade is A- on Pset 1 and 2

Midterm Feedback

- 20 responses out of a class of 44 (selection bias?)
- Lecture clear and engaging (8 people)
- Sections very helpful (4 people)
- Go slower when going through R in class (3 people)
- Pset expectations not clear (5 people)
  Median grade is A- on Pset 1 and 2
- Respondents divided on group activities:

Midterm Feedback

- 20 responses out of a class of 44 (selection bias?)
- Lecture clear and engaging (8 people)
- Sections very helpful (4 people)
- Go slower when going through R in class (3 people)
- Pset expectations not clear (5 people)
  Median grade is A- on Pset 1 and 2
- Respondents divided on group activities:
  - Group work is what's going well about class and there should be more (3 people)

Midterm Feedback

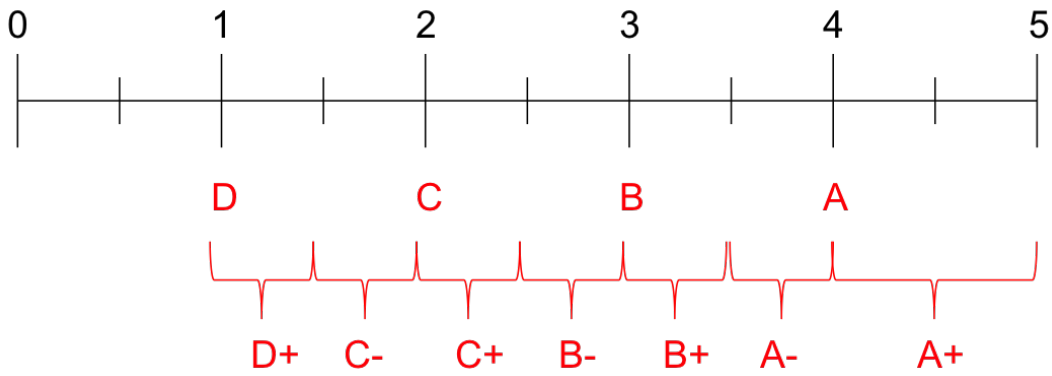
- 20 responses out of a class of 44 (selection bias?)
- Lecture clear and engaging (8 people)
- Sections very helpful (4 people)
- Go slower when going through R in class (3 people)
- Pset expectations not clear (5 people)
  Median grade is A- on Pset 1 and 2
- Respondents divided on group activities:
  - Group work is what's going well about class and there should be more (3 people)
  - Group work should be eliminated (2 people)

Grade Distribution

- A+: > 4
- A: 4
- A-: $>= 3.5$ to < 4
- B+: > 3 to < 3.5
- B: 3
- B-: $>= 2.5$ to < 3
- C+: > 2 to < 2.5
- C: 2
- C-: $>= 1.5$ to < 2

Grade Distribution

- A+: > 4
- A: 4
- A-: $>= 3.5$ to < 4
- B+: > 3 to < 3.5
- B: 3
- B-: $>= 2.5$ to < 3
- C+: > 2 to < 2.5
- C: 2
- C-: $>= 1.5$ to < 2



Where we are

Where we are

- Previously

Where we are

- Previously
  - Describing data

Where we are

- Previously
  - Describing data
  - Assessing hypotheses by making comparisons

Where we are

- Previously
  - Describing data
  - Assessing hypotheses by making comparisons
- This week

Where we are

- Previously
  - Describing data
  - Assessing hypotheses by making comparisons
- This week
  - Making comparisons with interval variables (correlation)

Where we are

- Previously
  - Describing data
  - Assessing hypotheses by making comparisons
- This week
  - Making comparisons with interval variables (correlation)
  - Thinking systematically about how X relates to Y (linear regression)

Making Comparisons

X	Y	How to Compare

Making Comparisons

X	Y	How to Compare
nominal	nominal	

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	cross-tabulation

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	cross-tabulation
ordinal	ordinal	

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	cross-tabulation
ordinal	ordinal	cross-tabulation

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	cross-tabulation
ordinal	ordinal	cross-tabulation
ordinal	interval	

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	cross-tabulation
ordinal	ordinal	cross-tabulation
ordinal	interval	comparison of means

Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	cross-tabulation
ordinal	ordinal	cross-tabulation
ordinal	interval	comparison of means
interval	interval	

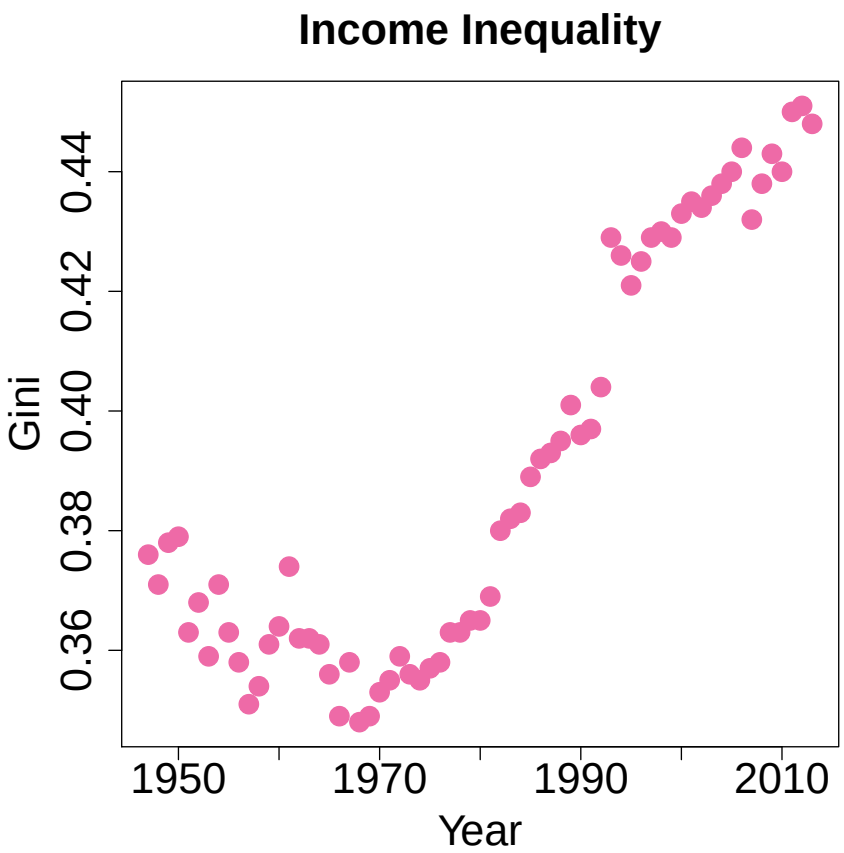
Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	cross-tabulation
ordinal	ordinal	cross-tabulation
ordinal	interval	comparison of means
interval	interval	`plot()`

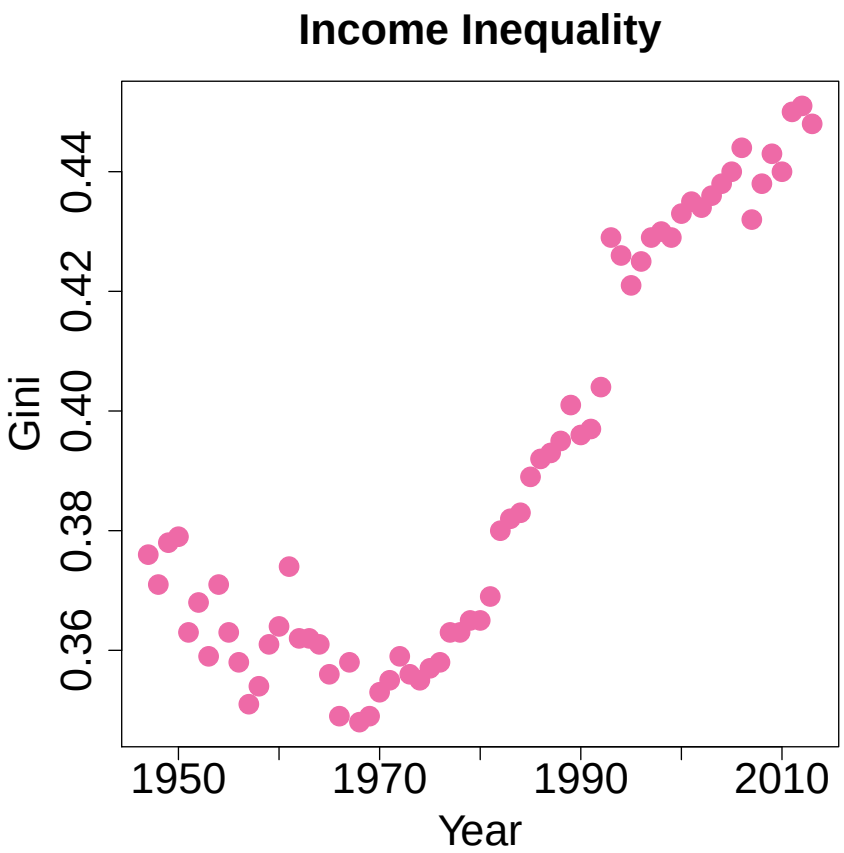
Making Comparisons

X	Y	How to Compare
nominal	nominal	cross-tabulation
nominal	ordinal	cross-tabulation
nominal	interval	comparison of means
ordinal	nominal	cross-tabulation
ordinal	ordinal	cross-tabulation
ordinal	interval	comparison of means
interval	interval	correlation

Scatterplot: `plot()`

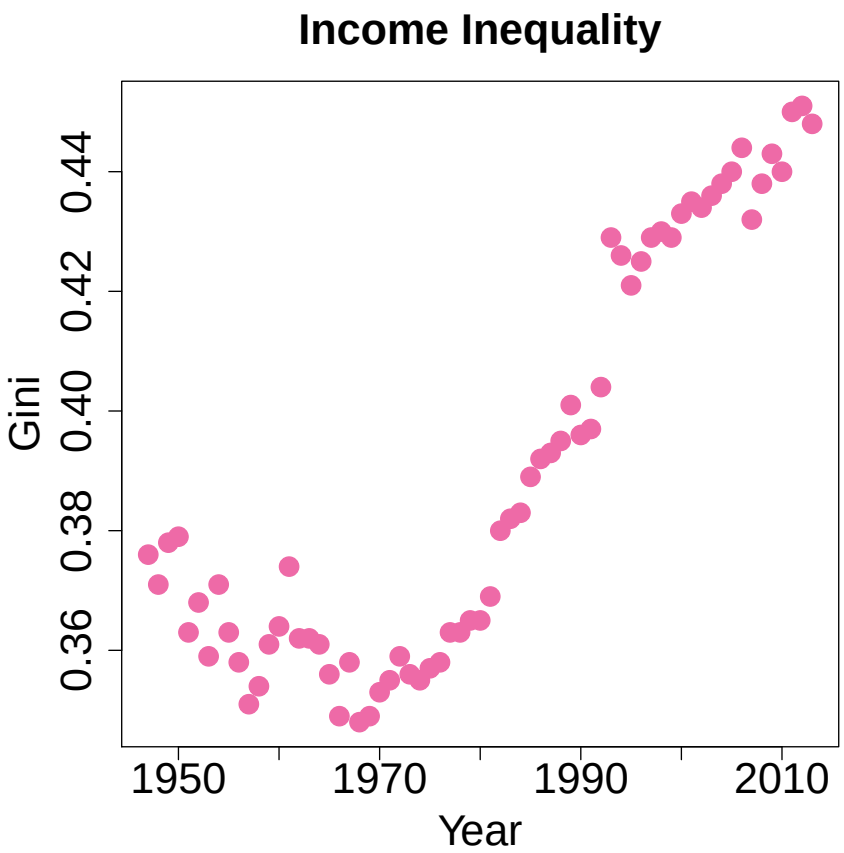


Scatterplot: `plot()`



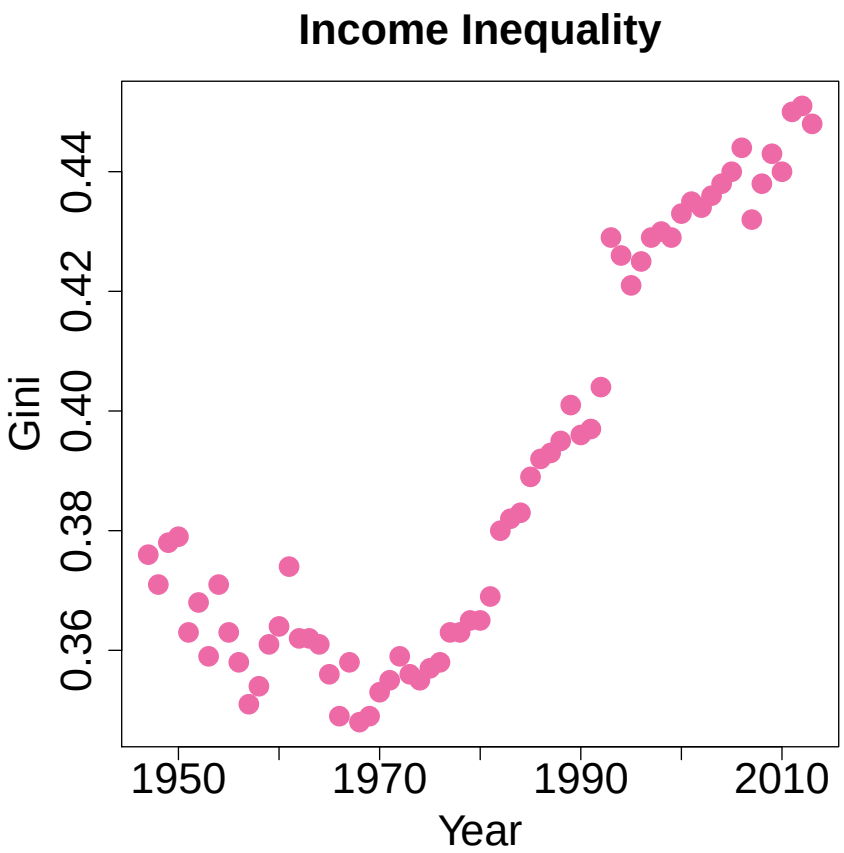
- For each year, the gini coefficient for the US

Scatterplot: `plot()`



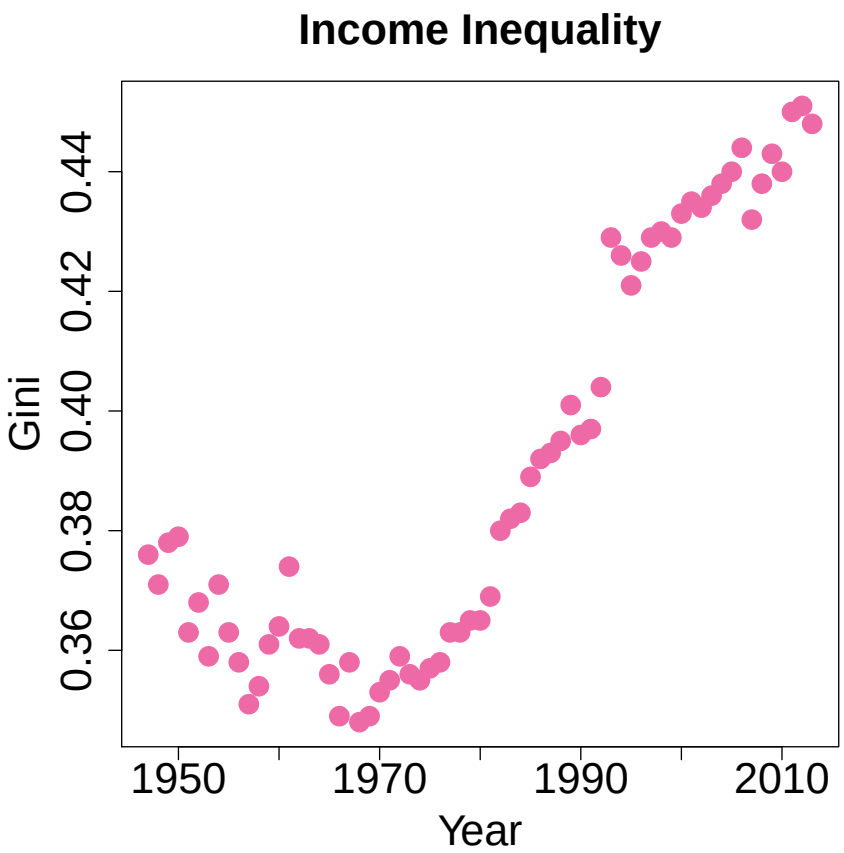
- For each year, the gini coefficient for the US
- Each point is an observation

Scatterplot: `plot()`



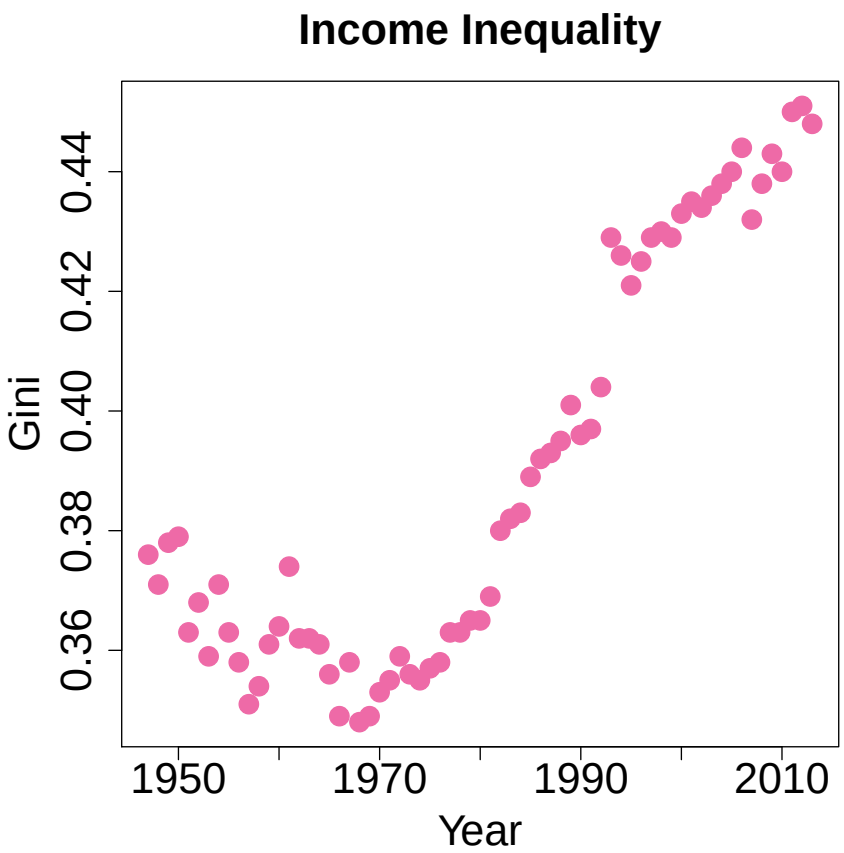
- For each year, the gini coefficient for the US
- Each point is an observation
- Each axis is a variable

Scatterplot: `plot()`



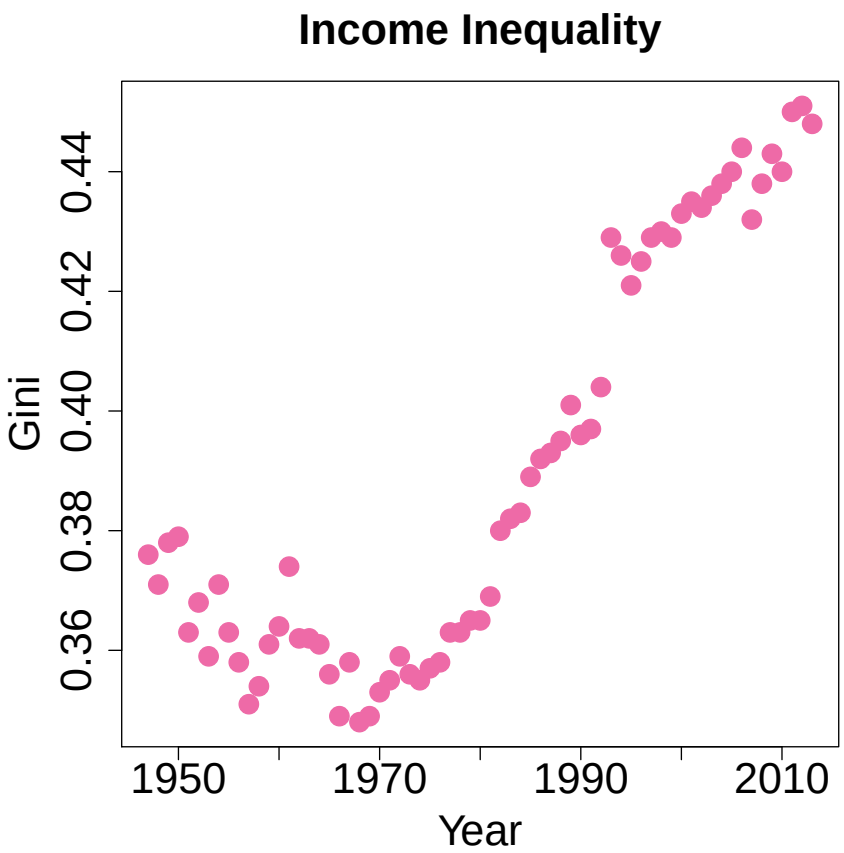
- For each year, the gini coefficient for the US
- Each point is an observation
- Each axis is a variable
- If you have an independent variable, it goes on the x-axis

Scatterplot: `plot()`



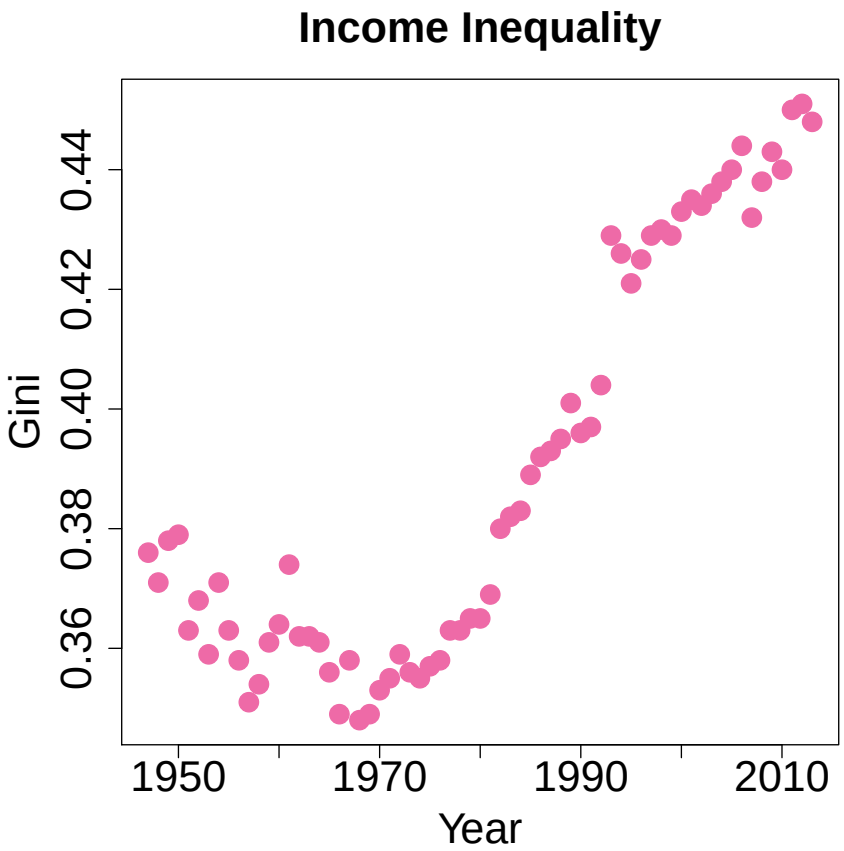
- For each year, the gini coefficient for the US
- Each point is an observation
- Each axis is a variable
- If you have an independent variable, it goes on the x-axis
- Dependent on y-axis

Scatterplot: `plot()`



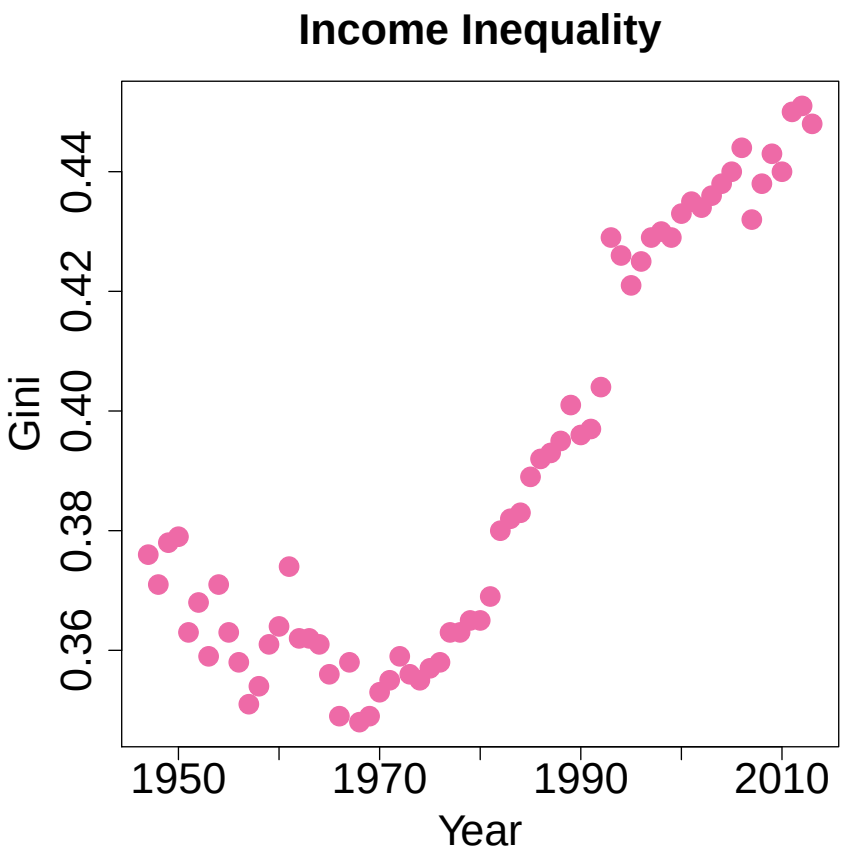
- For each year, the gini coefficient for the US
- Each point is an observation
- Each axis is a variable
- If you have an independent variable, it goes on the x-axis
- Dependent on y-axis
- Great 'first look' at your data

Scatterplot: So...



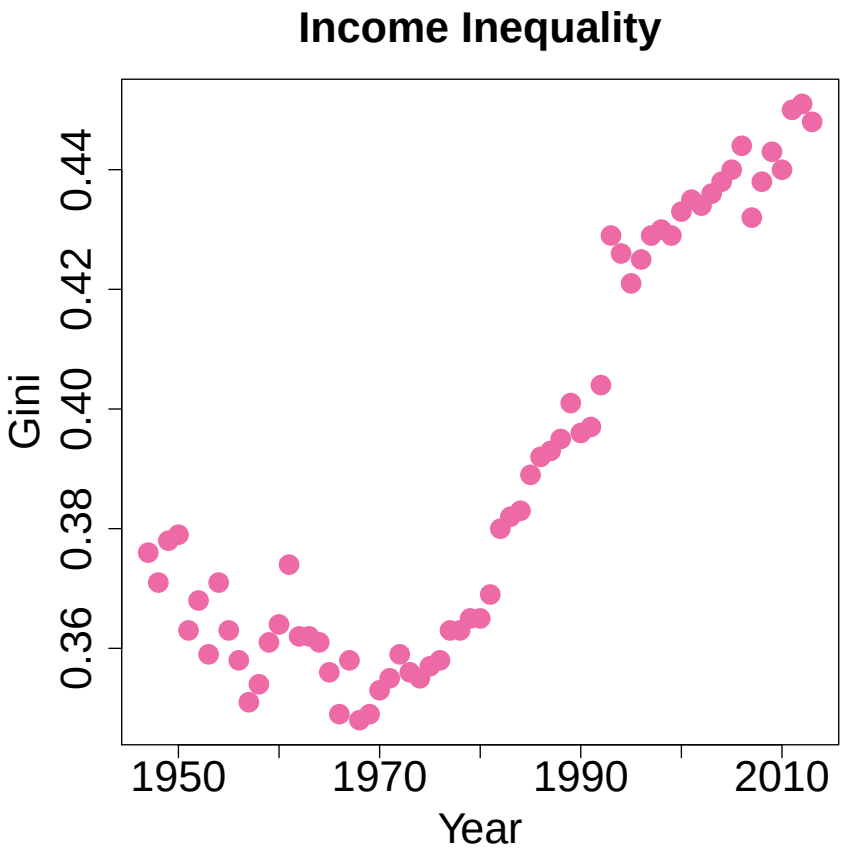
- Pattern:
- Direction:
- Strength:

Scatterplot: So...



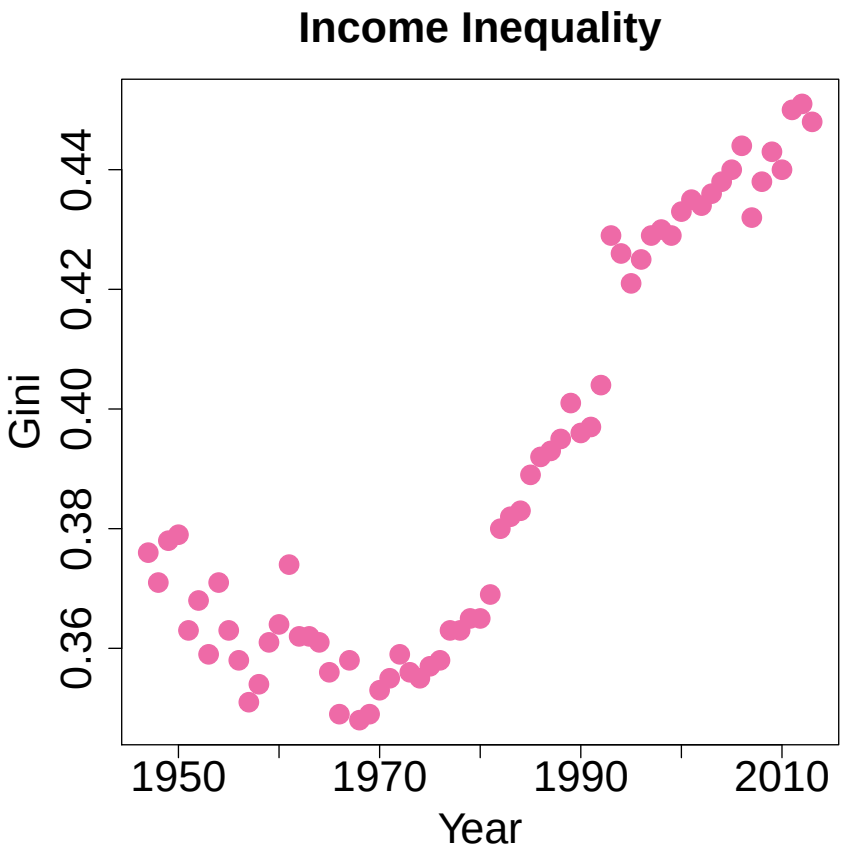
- Pattern: curvilinear
- Direction:
- Strength:

Scatterplot: So...



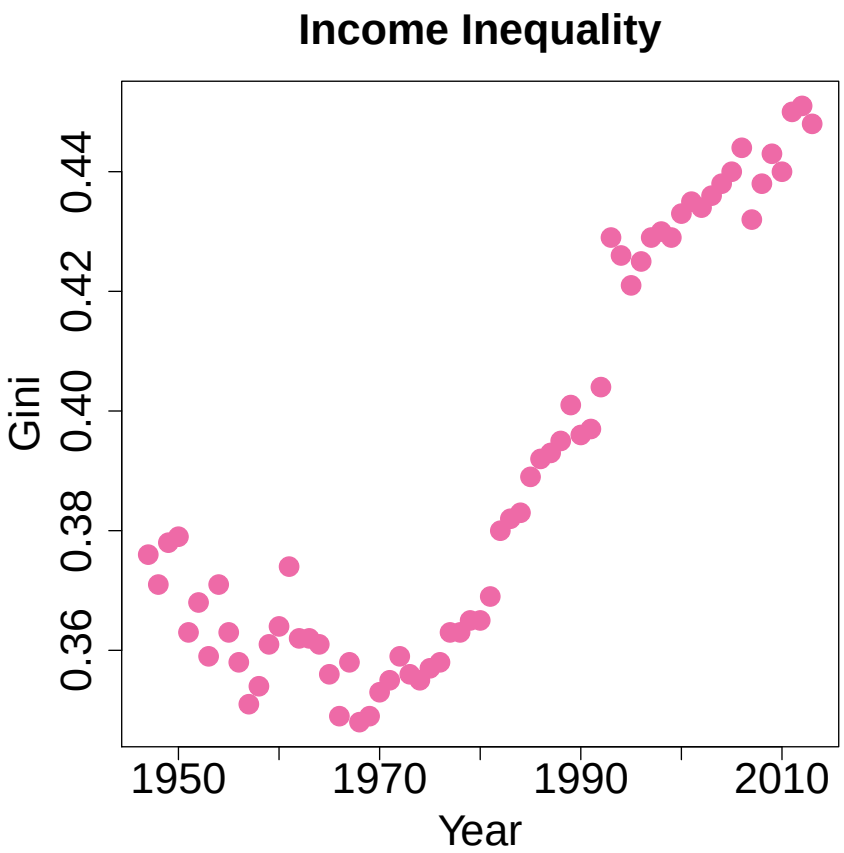
- Pattern: curvilinear
- Direction: negative, then positive; income inequality in the US declined until the 1970s, then increased
- Strength:

Scatterplot: So...



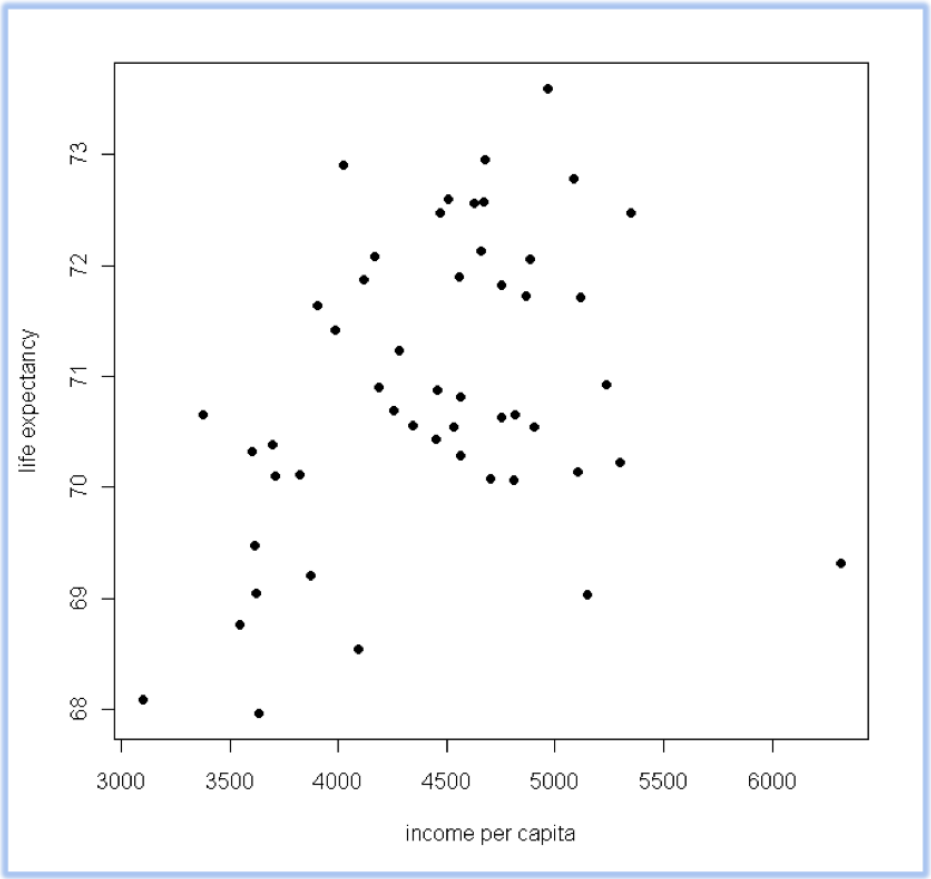
- Pattern: curvilinear
- Direction: negative, then positive; income inequality in the US declined until the 1970s, then increased
- Strength:
  - Pretty consistent

Scatterplot: So...



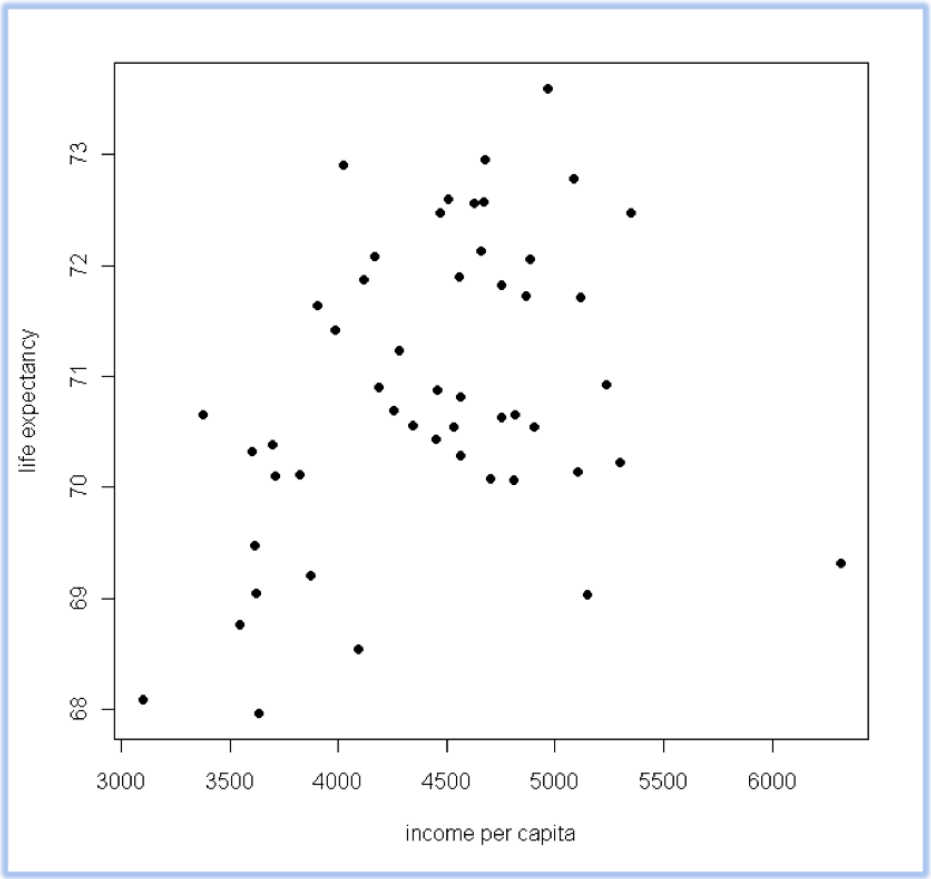
- Pattern: curvilinear
- Direction: negative, then positive; income inequality in the US declined until the 1970s, then increased
- Strength:
  - Pretty consistent
  - Very few exceptions

Scatterplot: Another one



- Pattern:
- Direction:
- Strength:
  - Consistency:
  - Exceptions:

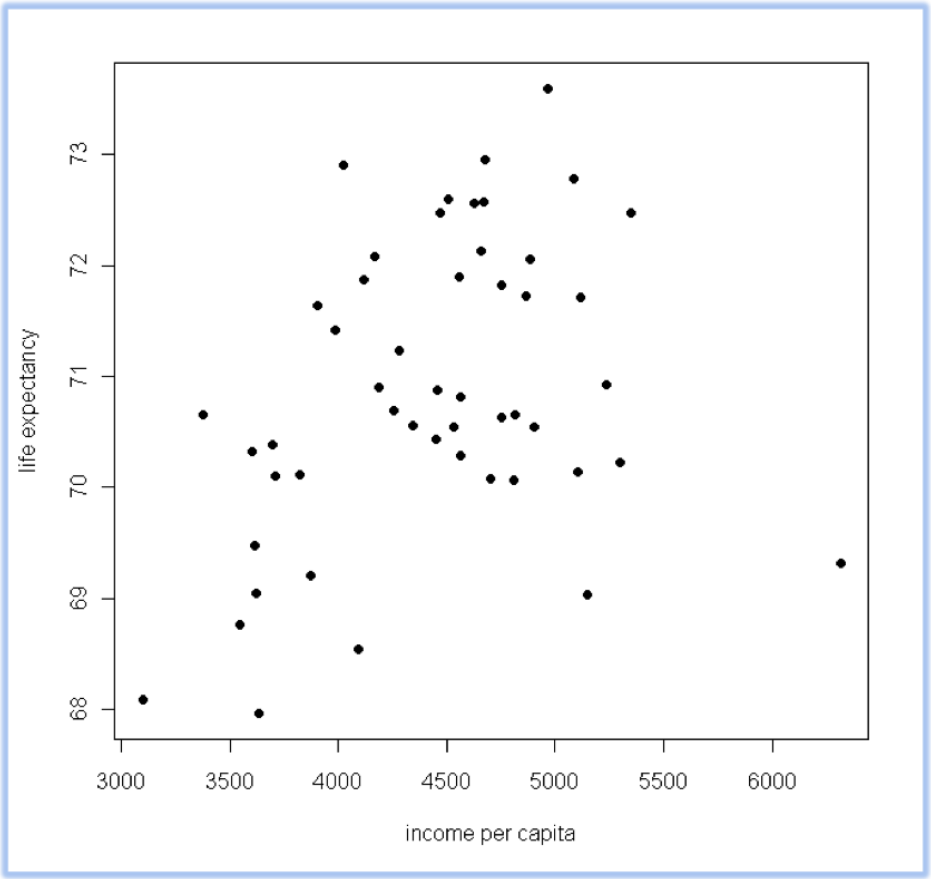
Scatterplot: Another one



- Pattern: linear?
- Direction:

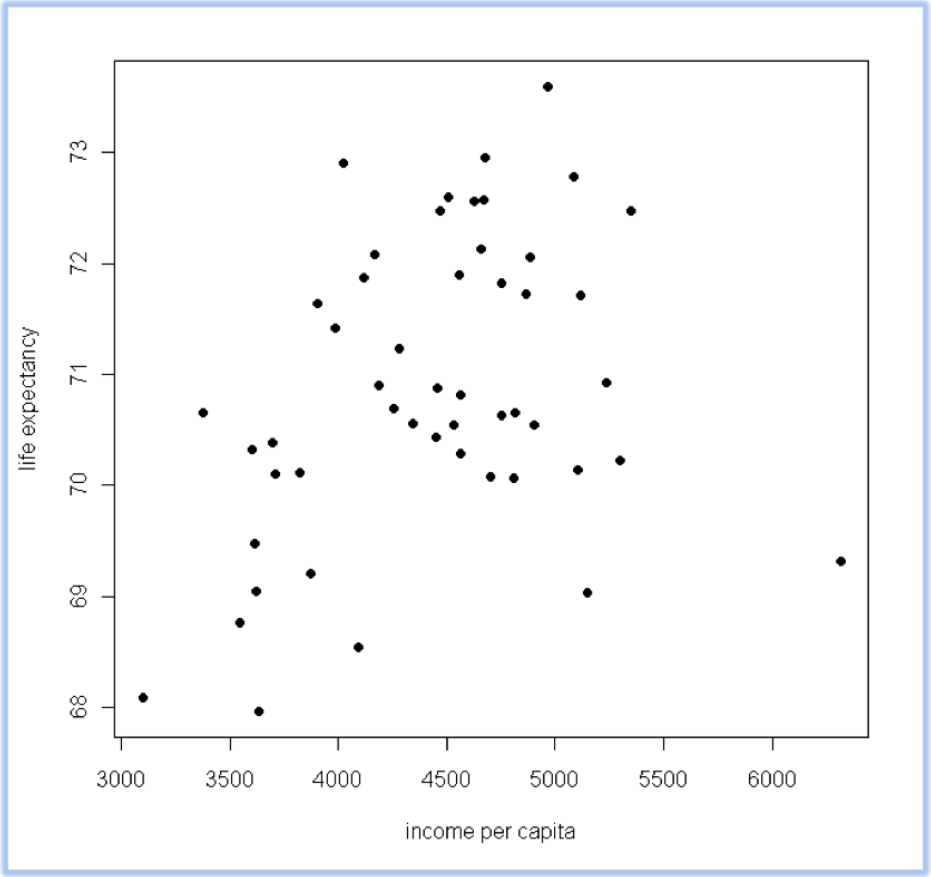
- Strength:
  - Consistency:
  - Exceptions:

Scatterplot: Another one



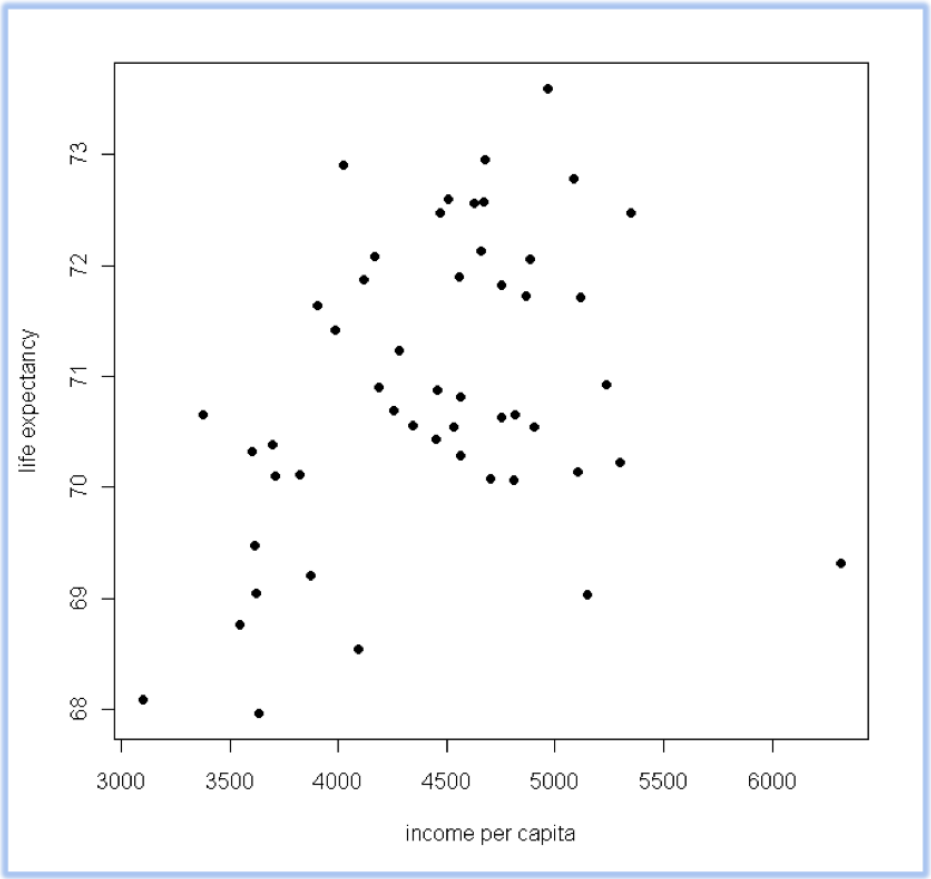
- Pattern: linear?
- Direction: positive; as income increases people generally live longer
- Strength:
  - Consistency:
  - Exceptions:

Scatterplot: Another one



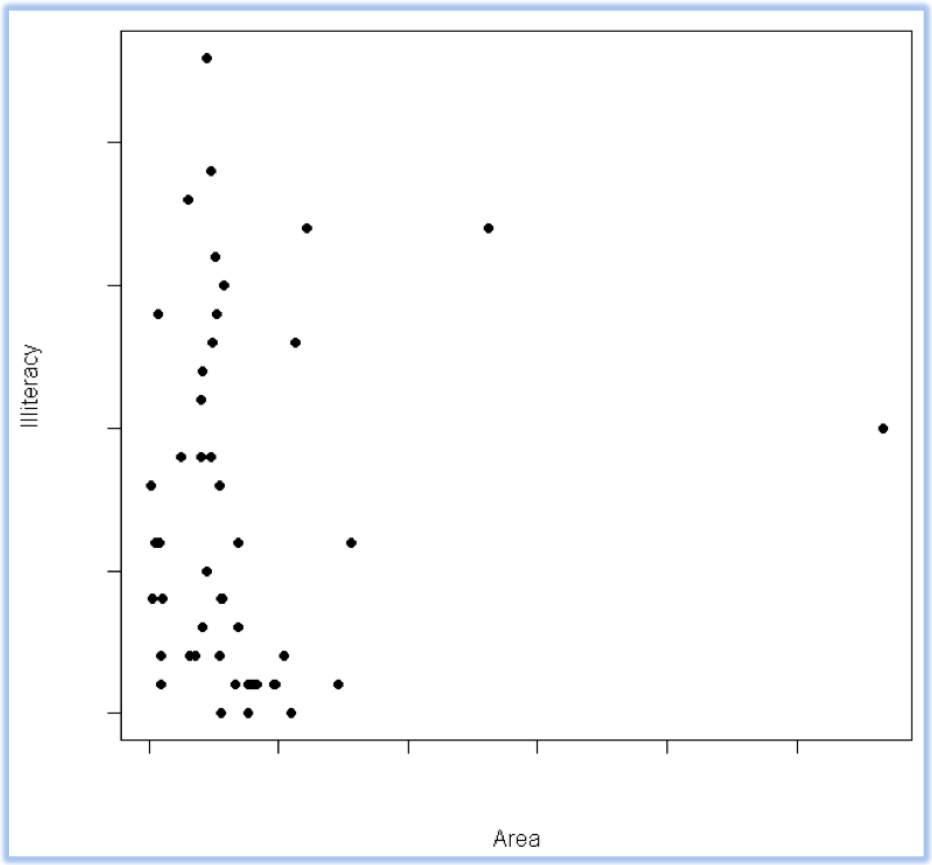
- Pattern: linear?
- Direction: positive; as income increases people generally live longer
- Strength:
  - Consistency: moderate
  - Exceptions:

Scatterplot: Another one



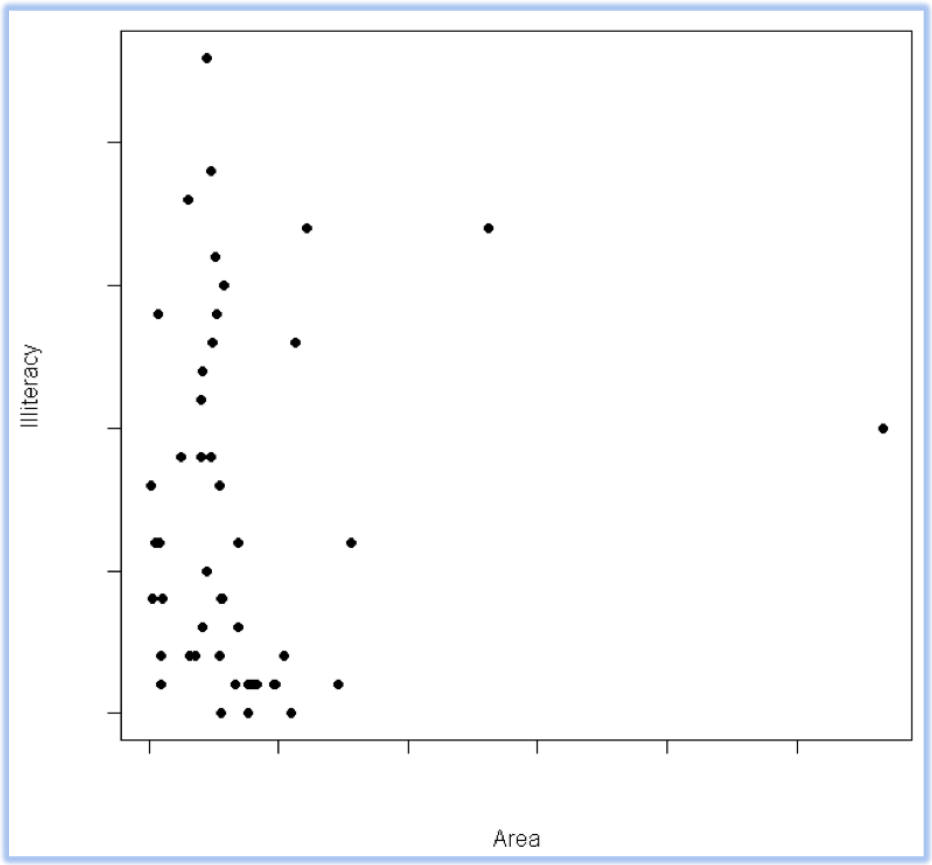
- Pattern: linear?
- Direction: positive; as income increases people generally live longer
- Strength:
  - Consistency: moderate
  - Exceptions: some exceptions to trend!

Scatterplot: One more



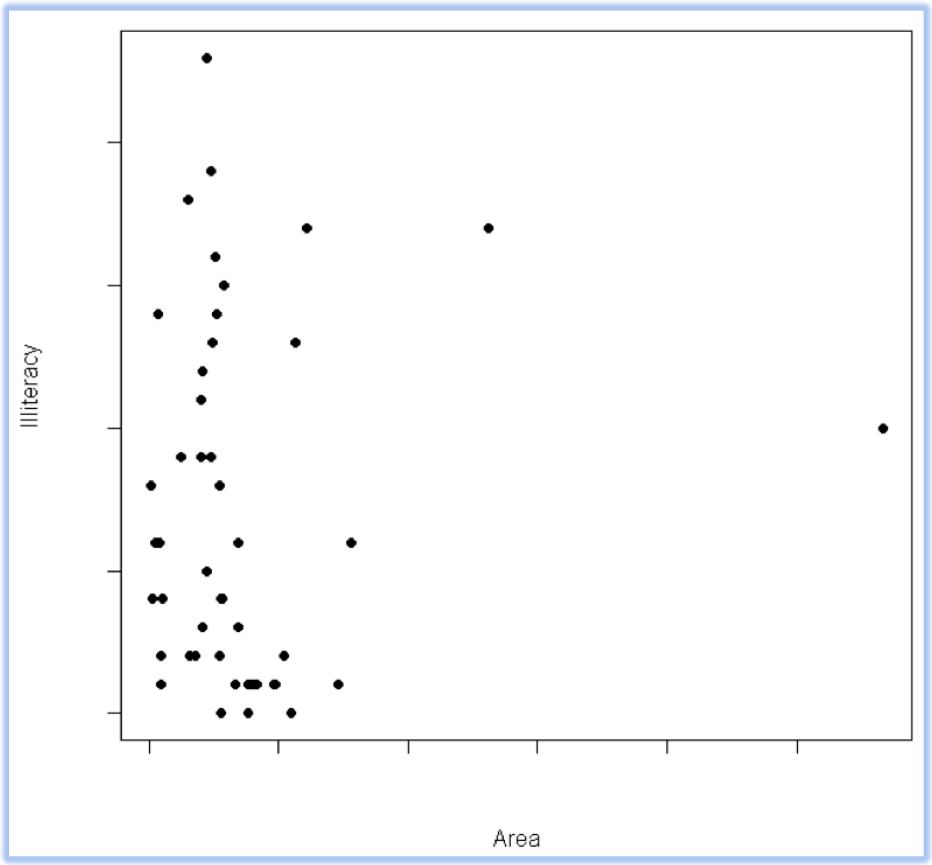
- Pattern:
- Direction:
- Strength:
  - Consistency:
  - Exceptions:

Scatterplot: One more



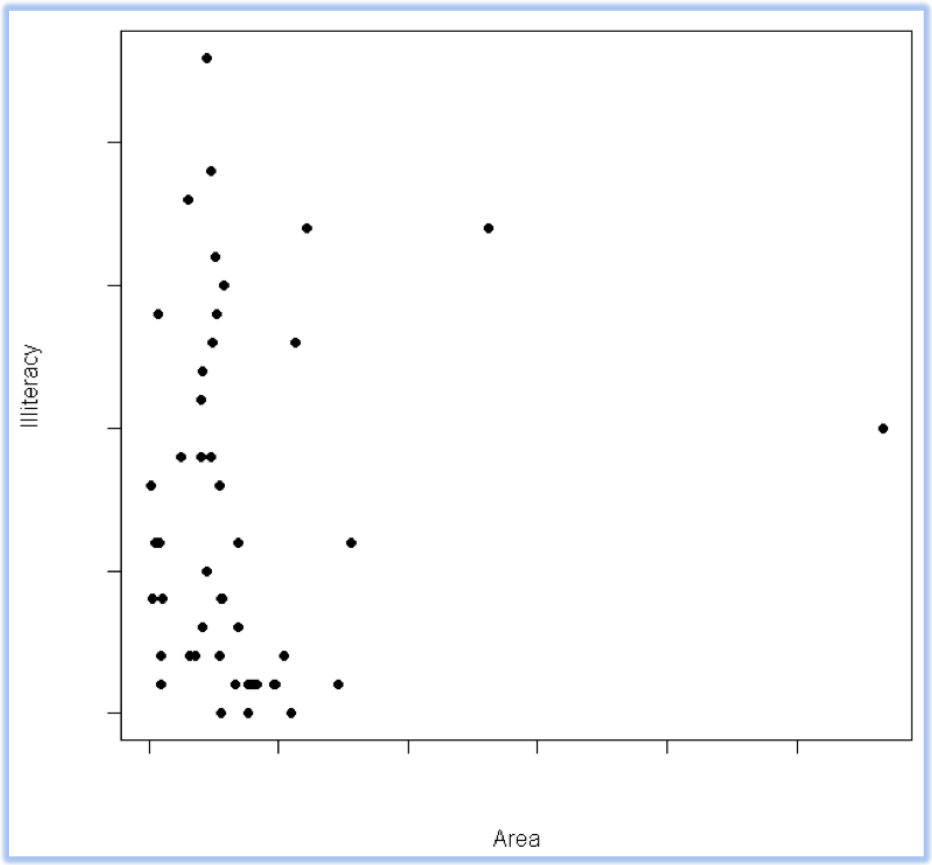
- Pattern: not obvious
- Direction:
- Strength:
  - Consistency:
  - Exceptions:

Scatterplot: One more



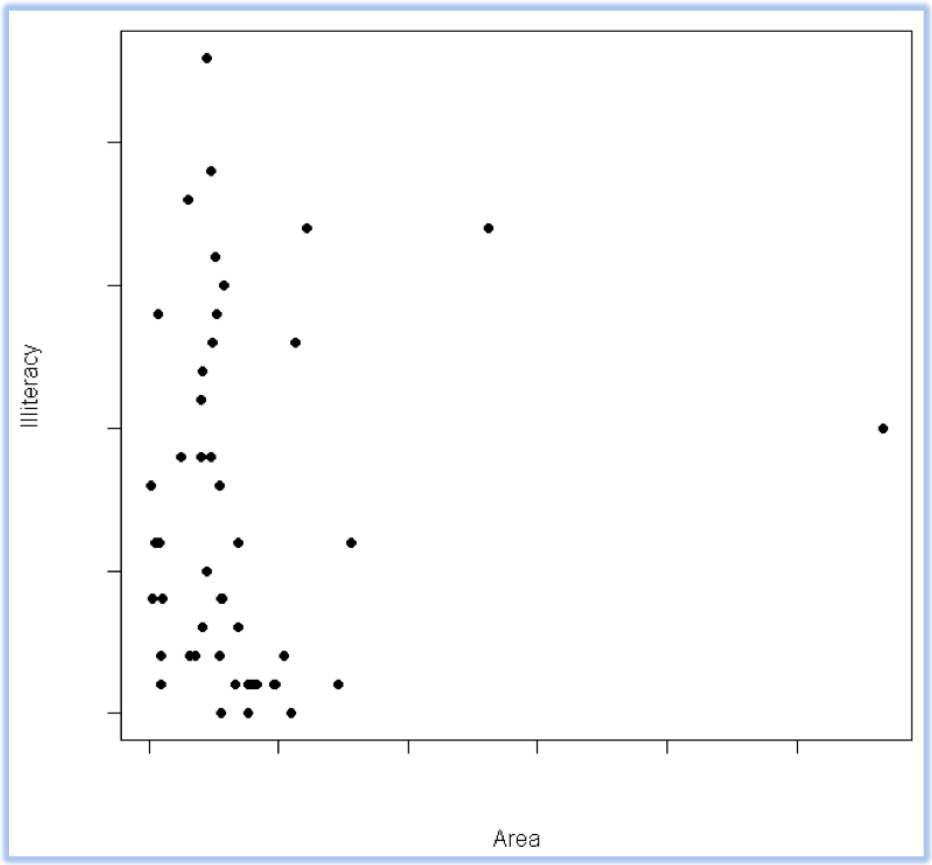
- Pattern: not obvious
- Direction: positive, maybe?
- Strength:
  - Consistency:
  - Exceptions:

Scatterplot: One more



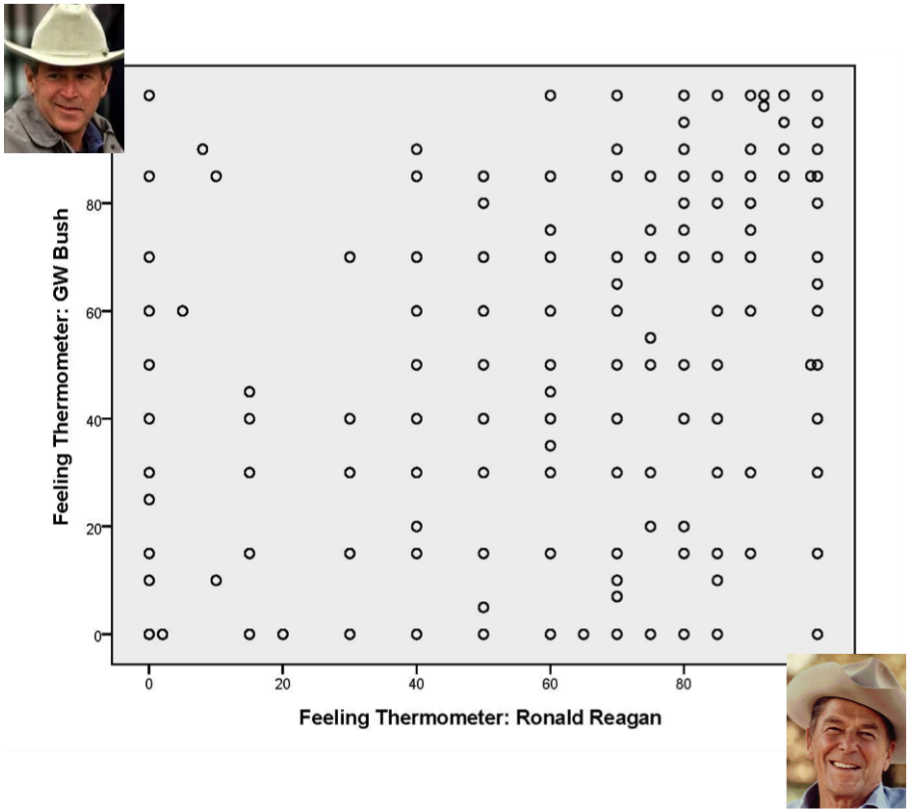
- Pattern: not obvious
- Direction: positive, maybe?
- Strength:
  - Consistency: very week
  - Exceptions:

Scatterplot: One more



- Pattern: not obvious
- Direction: positive, maybe?
- Strength:
  - Consistency: very week
  - Exceptions: no obvious relationship

Scatterplot: Be Careful



- Sometimes observations stack up on same point
- Can hide general pattern (looks like no relationship, but actually a moderately strong relationship!)
- Happens when individuals approximate responses (round up or round down)

Making Comparisons with Interval Data

an objective measure...

...for direction, and strength

Pearson correlation coefficient

is *r* in the *sample*...

"rho" said "row"

...ρ in *population*

"how closely do the data follow a (straight line) trend?"

"how close do data cluster around a linear pattern?"

Making Comparisons with Interval Data: Features of r

- larger absolute value implies stronger (linear) association
- can take any value between...

% vaccinated and % at risk from disease

wing length and beats per min

height of twins

-1 is *perfect* negative (linear) association

and

1 is *perfect* positive (linear) association

men's shoe size and foot length

0 implies **no** (linear) association between the variables

Making Comparisons with Interval Data: Notice

- larger absolute value implies stronger (linear) association
- can take any value between...

% vaccinated and % at risk from disease

wing length and beats per min

height of twins

-1 is *perfect* negative (linear) association

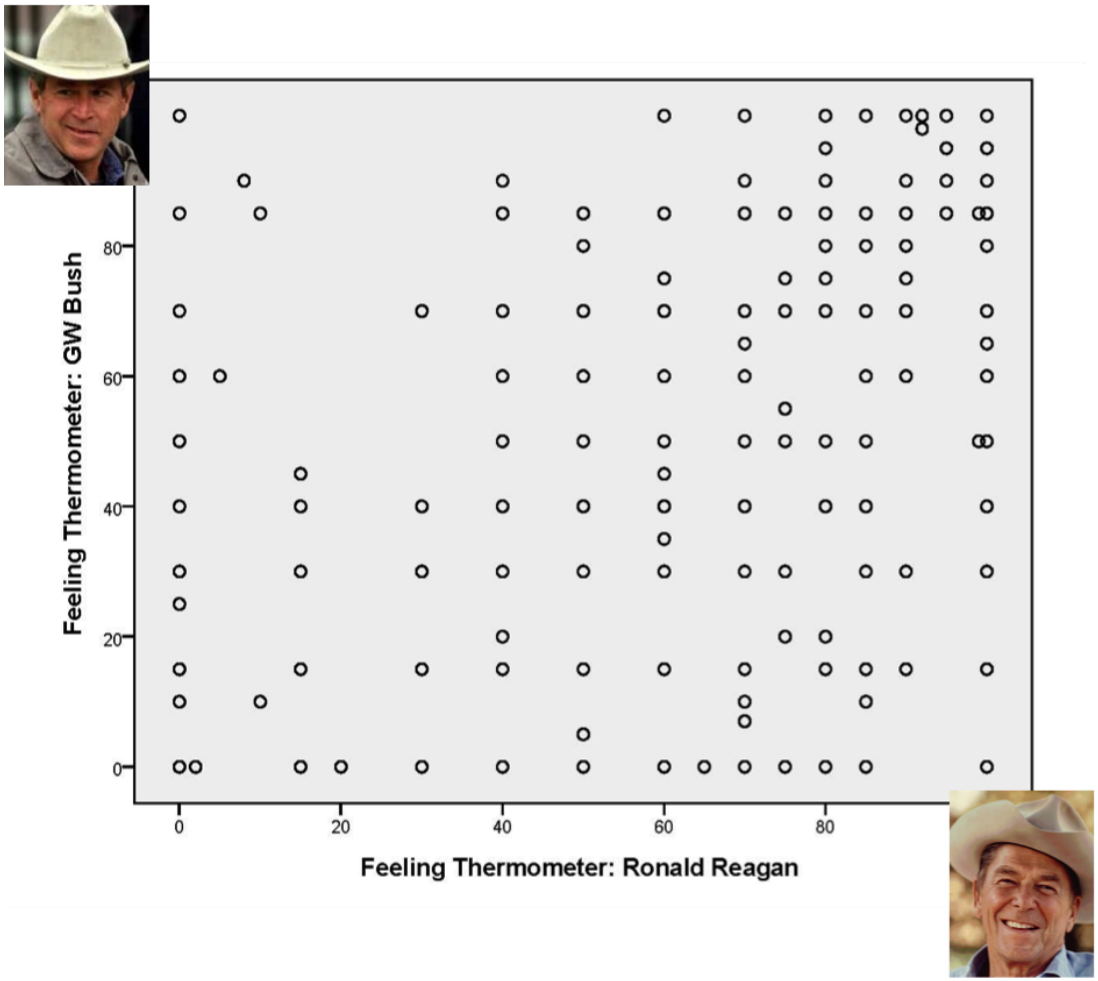
and

1 is *perfect* positive (linear) association

men's shoe size and foot length

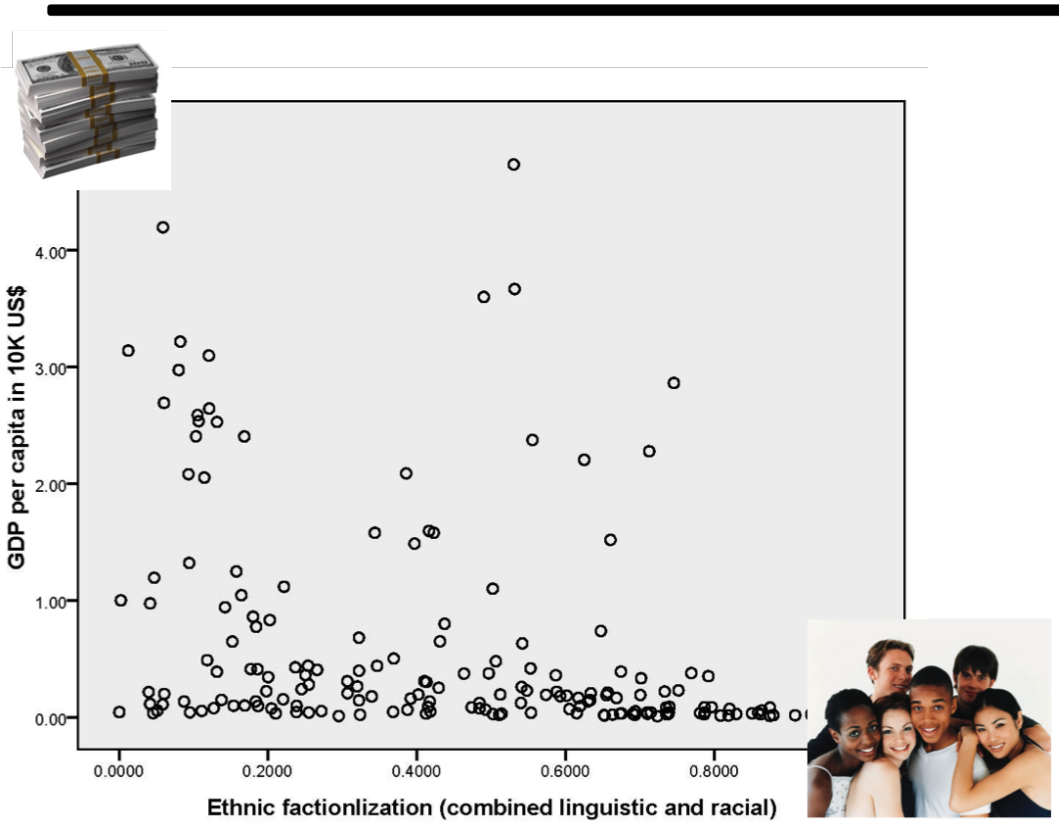
0 implies **no** (linear) association between the variables

Scatterplot vs Correlation



- $r = 0.589$ (linear?)
- moderately strong, positive ($r > 0$) relationship

Scatterplot vs Correlation



- $r = -0.321$ (linear?)
- weak, negative ($r < 0$) relationship

Scatterplot vs Correlation

- $r = 0.810$ (close to linear)
- strong, positive ($r > 0$) relationship

Correlation Formula

$$\frac{\sum\left(\frac{X_i - \overline{X}}{s_x}\right)\left(\frac{Y_i - \overline{Y}}{s_y}\right)}{n-1}$$

Correlation Formula

$$\frac{\sum\left(\frac{X_i - \overline{X}}{s_x}\right)\left(\frac{Y_i - \overline{Y}}{s_y}\right)}{n-1}$$

1. Standardize every observation of X: subtract from mean of X and divide by sd of X

Correlation Formula

$$\frac{\sum\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)}{n-1}$$

1. Standardize every observation of X: subtract from mean of X and divide by sd of X
2. Standardize every observation of Y: subtract from mean of Y and divide by sd of Y

Correlation Formula

$$\frac{\sum\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)}{n-1}$$

1. Standardize every observation of X: subtract from mean of X and divide by sd of X
2. Standardize every observation of Y: subtract from mean of Y and divide by sd of Y
3. Multiplying (1) and (2) together

Correlation Formula

$$\frac{\sum\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)}{n-1}$$

1. Standardize every observation of X: subtract from mean of X and divide by sd of X
2. Standardize every observation of Y: subtract from mean of Y and divide by sd of Y
3. Multiplying (1) and (2) together
4. Add up, then divide by sample size (N) minus one

Correlation Formula Example

Suppose x is

1,2,4,1

...then, xbar is 2, sd is 1.41

The Z-scores are

(1-2)/1.41, (2-2)/1.41,
(4-2)/1.41, (1-2)/1.41

or

-0.71, 0.00, 1.42, -0.71

Suppose y is

1,3,5,3

...then, xbar is 3, sd is 1.63

The Z-scores are

(1-3)/1.63, (3-3)/1.63,
(5-3)/1.63, (3-3)/1.63

or

-1.23, 0.00, 1.23, 0.00

	-0.71	0.00	1.42	-0.71
multiply	-1.23	0.00	1.23	0.00
	0.87	0.00	1.74	0.00

$r = (.87+0+1.74+0)/n-1$
$= 0.87$

Correlation: R to the Rescue

Correlation: R to the Rescue

- `cor()`

Correlation: R to the Rescue

- `cor()`
- US inequality and political polarization (difference in mean voting patterns of democrats and republicans in Congress)

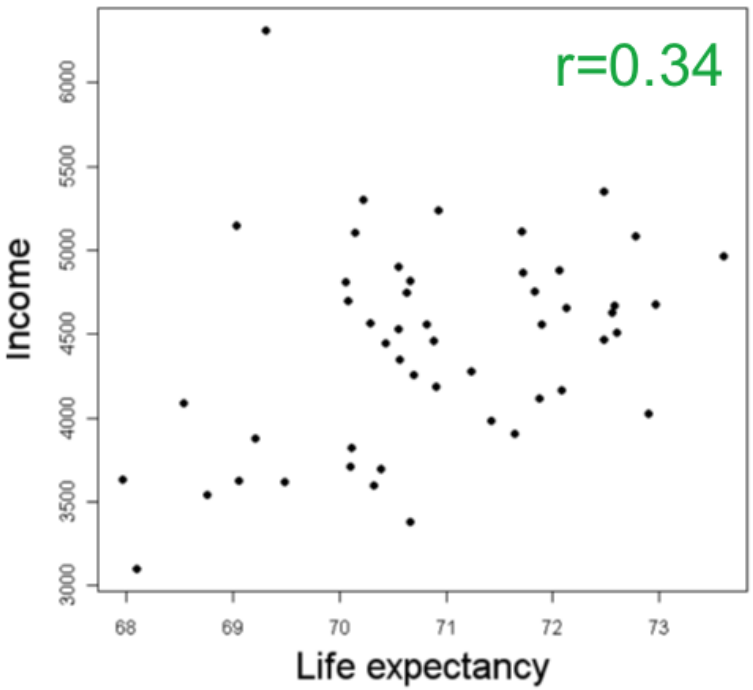
Correlation: Symmetry

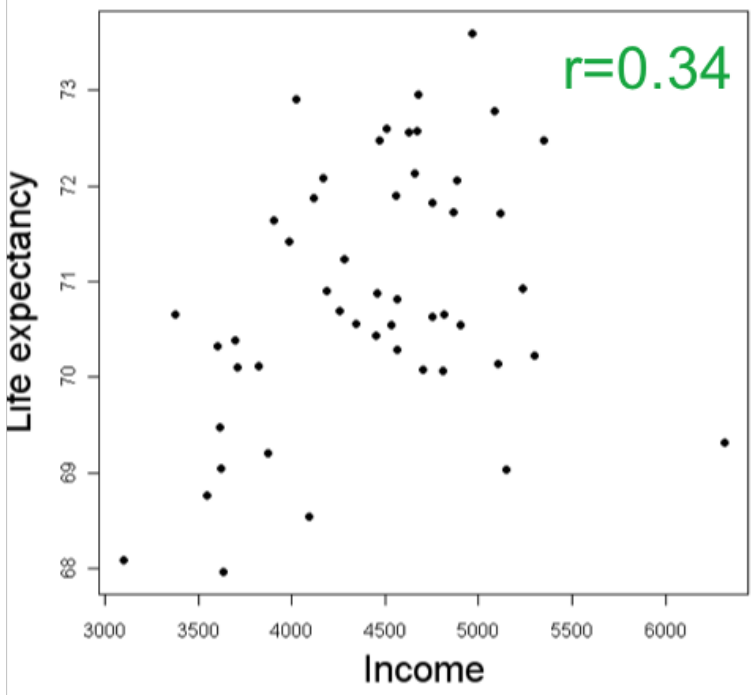
Correlation: Symmetry

- r is "symmetric"

Correlation: Symmetry

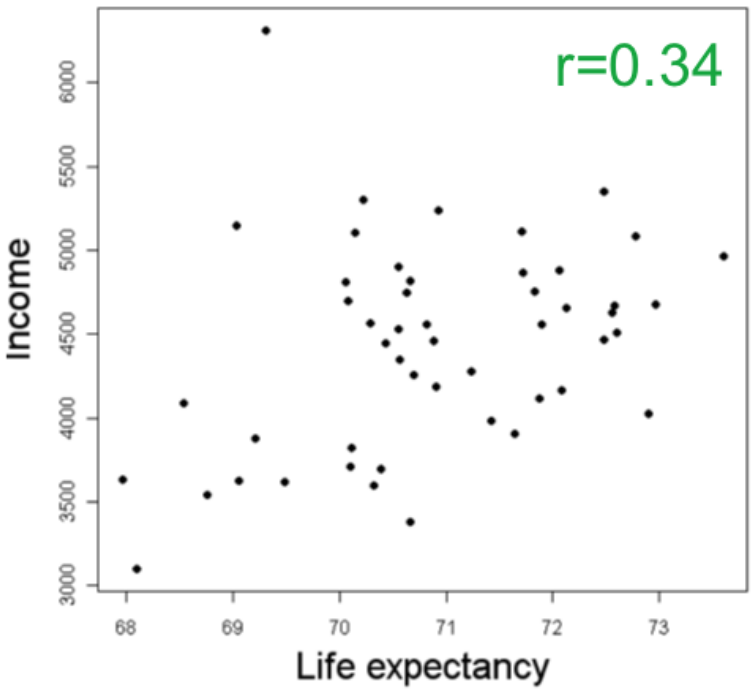
- r is "symmetric"
- Correlation between X and Y...is same as between Y and X

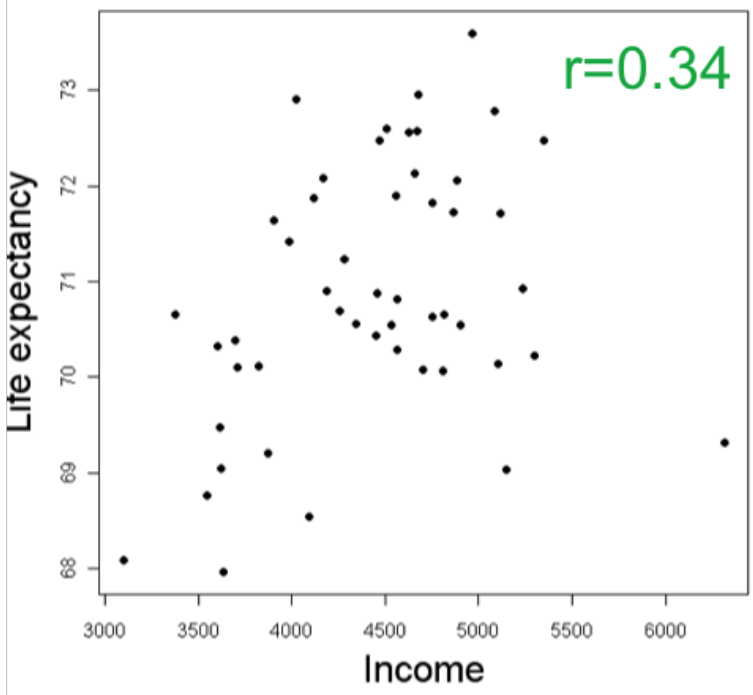




Correlation: Symmetry

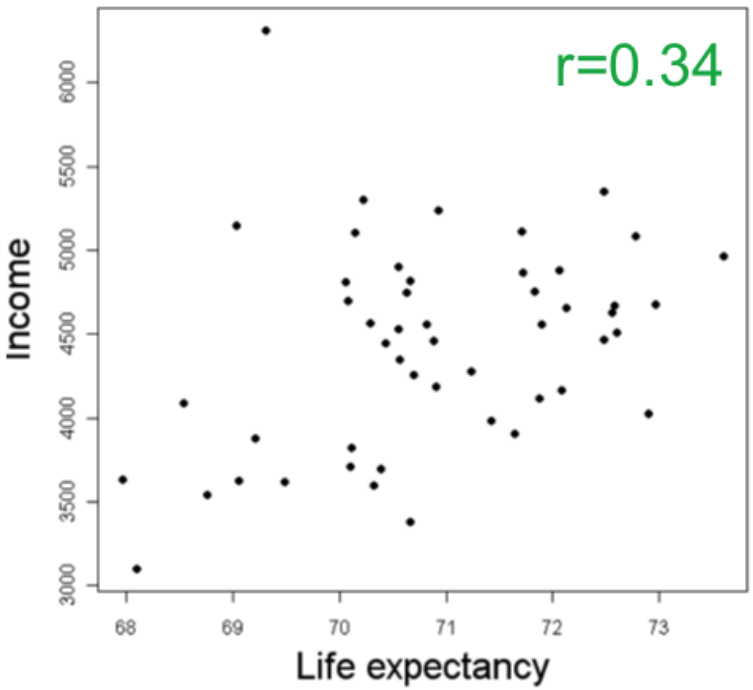
- r is "symmetric"
- Correlation between X and Y...is same as between Y and X
- Even if plots look different

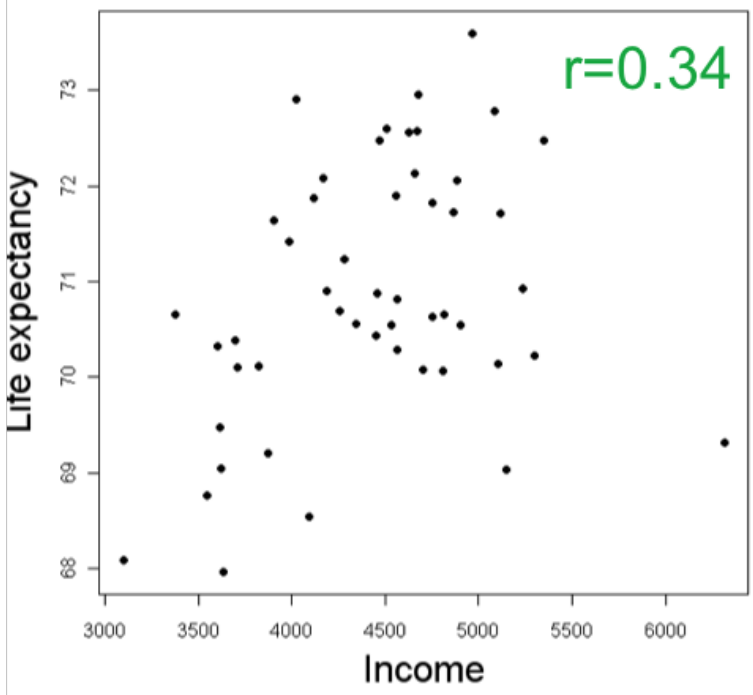




Correlation: Symmetry

- r is "symmetric"
- Correlation between X and Y...is same as between Y and X
- Even if plots look different
- Correlation doesn't change according to which variable is dependent or independent





Correlation: Unit-less

Correlation: Unit-less

- Correlation coefficient (r) is not expressed in terms of units of original variables

Correlation: Unit-less

- Correlation coefficient (r) is not expressed in terms of units of original variables
- Makes no difference what original units of variable were

Correlation: Unit-less

- Correlation coefficient (r) is not expressed in terms of units of original variables
- Makes no difference what original units of variable were

Example:

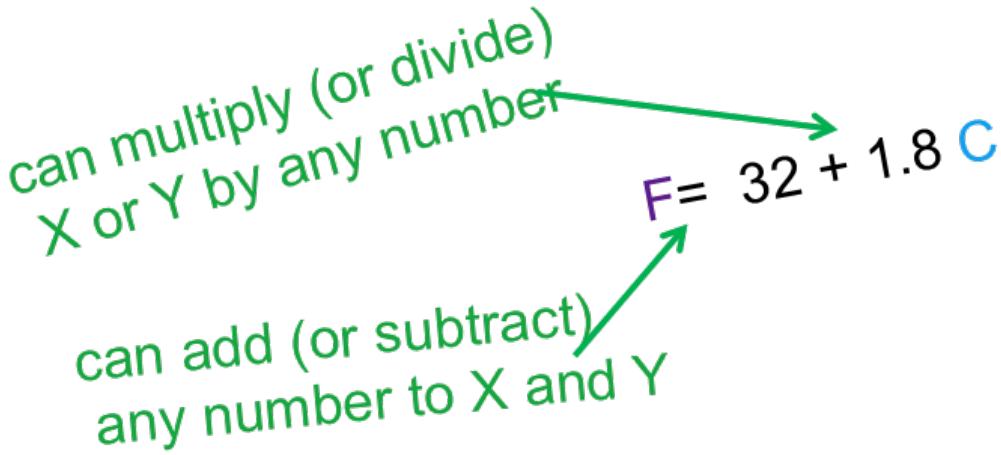
- Correlation between daily max temp in Palo Alto and Berkeley is 0.72...

Correlation: Unit-less

- Correlation coefficient (r) is not expressed in terms of units of original variables
- Makes no difference what original units of variable were

Example:

- Correlation between daily max temp in Palo Alto and Berkeley is 0.72...
- Can convert one city to degrees Celsius...or both to Celsius...or both to Fahrenheit (or Kelvin!)

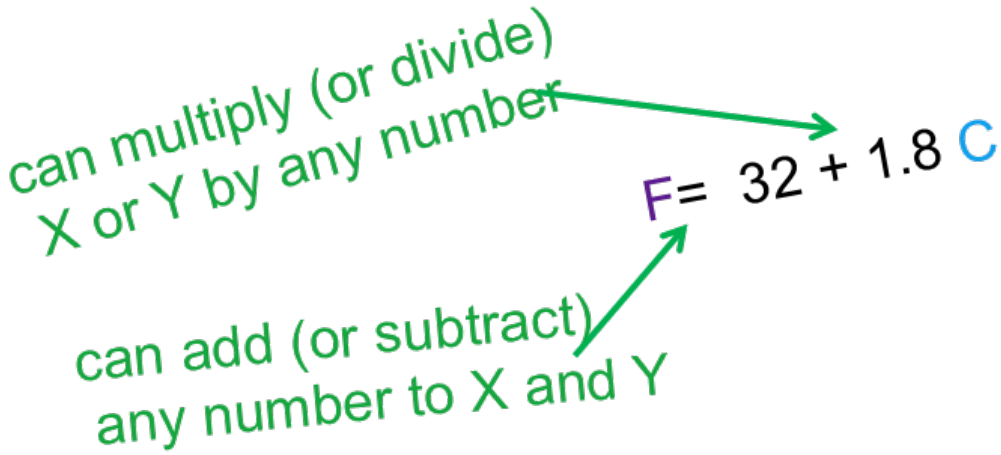


Correlation: Unit-less

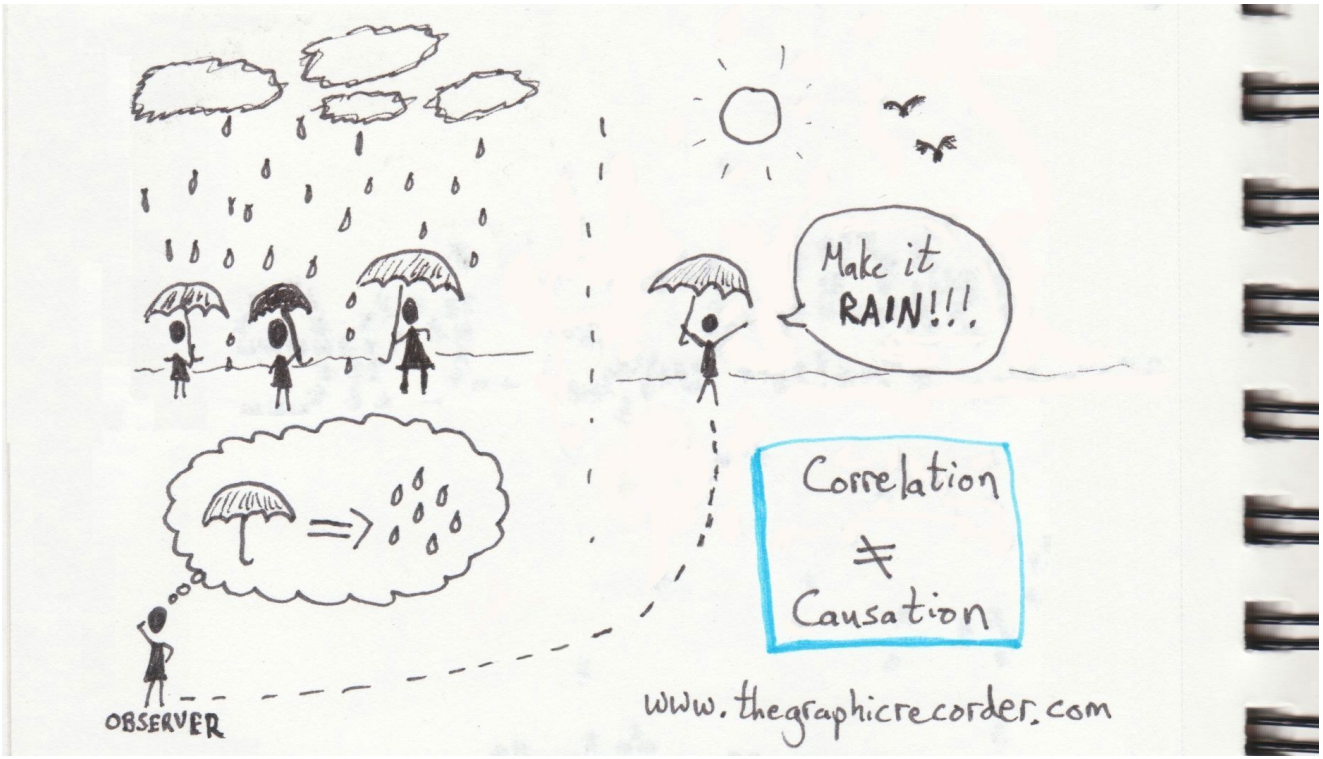
- Correlation coefficient (r) is not expressed in terms of units of original variables
- Makes no difference what original units of variable were

Example:

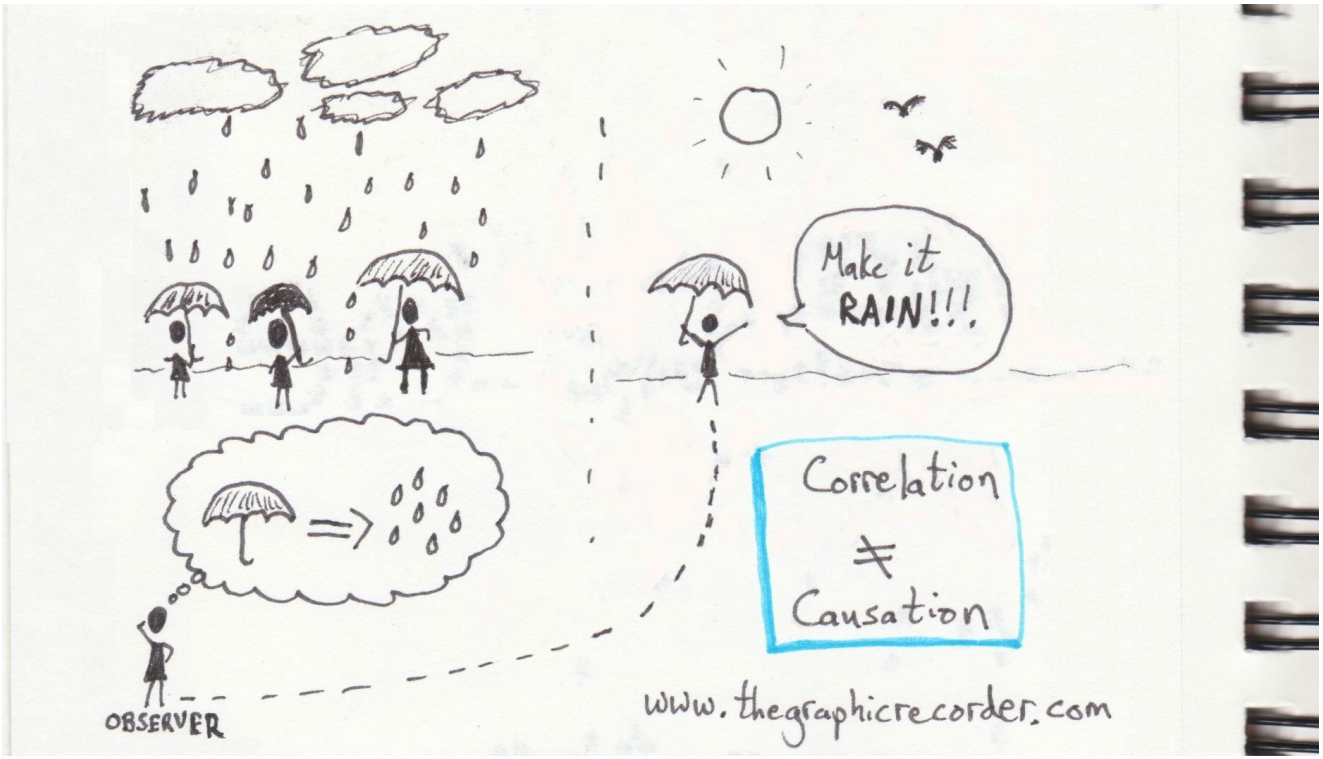
- Correlation between daily max temp in Palo Alto and Berkeley is 0.72...
- Can convert one city to degrees Celsius...or both to Celsius...or both to Fahrenheit (or Kelvin!)
- r remains 0.72!



Correlation Does not Imply Causation

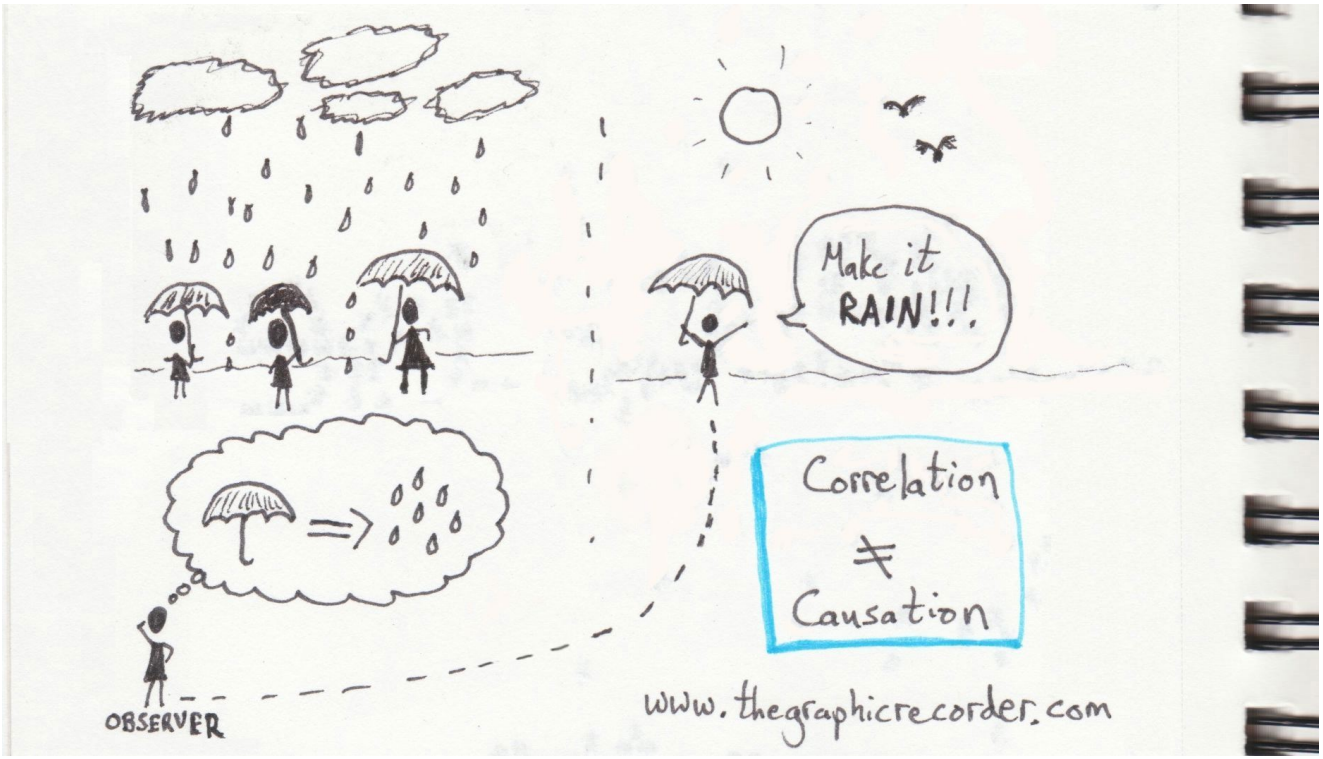


Correlation Does not Imply Causation



- If X and Y are correlated in the sample

Correlation Does not Imply Causation



- If X and Y are correlated in the sample
- X may cause Y...Y may cause X...Z may cause both...pure chance (random sampling error)

Examining Your Data

Examining Your Data

- Goal: Practice working with real data in R; make the concepts we've been learning in class application to your work and life

Examining Your Data

- Goal: Practice working with real data in R; make the concepts we've been learning in class application to your work and life
- Work in groups (learn from your peers)

Examining Your Data

- Goal: Practice working with real data in R; make the concepts we've been learning in class application to your work and life
- Work in groups (learn from your peers)
- Decided on a dataset (your own, one you've come across)

Examining Your Data

- Goal: Practice working with real data in R; make the concepts we've been learning in class application to your work and life
- Work in groups (learn from your peers)
- Decided on a dataset (your own, one you've come across)
- Bring that dataset to class next Wed 2/24 (csv from excel or google sheets)

Examining Your Data

- Goal: Practice working with real data in R; make the concepts we've been learning in class application to your work and life
- Work in groups (learn from your peers)
- Decided on a dataset (your own, one you've come across)
- Bring that dataset to class next Wed 2/24 (csv from excel or google sheets)
- Explore that data in class next Wed 2/24 (plots, descriptive stats)

Examining Your Data

- Goal: Practice working with real data in R; make the concepts we've been learning in class application to your work and life
- Work in groups (learn from your peers)
- Decided on a dataset (your own, one you've come across)
- Bring that dataset to class next Wed 2/24 (csv from excel or google sheets)
- Explore that data in class next Wed 2/24 (plots, descriptive stats)
- Share