Section 3: Relations between multiple variables

STOR 155.02, Spring '21

updated 2021-01-30

What you will learn

- time series (slides are your best reference!)
- scatterplots
- correlation, covariance

Resources

- Textbook ch 1.6 (scatterplots), ch 2.1.4 (correlation, linear relationships)
- Warning: Textbook does not have a dedicated section on time series, though they do appear. These notes are your main resource.

Time series

A time series is a numeric variable

with a special relationship to another variable,

TIME

measured in some units,

e.g. seconds, months, whatever

Show up *everywhere*

gross domestic product, arctic sea ice extent, your college grades, K-pop group sizes, crude birth rate, average daily temperature, employment rate, covid cases/deaths, police shootings, concentration of a contaminant in river water, song popularity, movie studio revenues, presidential approval ratings and so so much baseball data

Today's example: Female life expectancy at birth, World Bank data

annual, from 1960-2018, also on the course github page.

year	Azerbaijan	China	Ireland	Mali	Thailand	United.States	Zimbabwe
1960	64.168	45.191	71.707	29.026	57.059	73.1	54.672
1961	64.403	45.497	71.941	29.210	57.617	73.6	55.141
1962	64.629	46.243	72.147	29.432	58.151	73.5	55.609
1963	64.852	47.486	72.334	29.704	58.653	73.4	56.071
1964	65.074	49.194	72.510	30.033	59.120	73.7	56.521
1965	65.299	51.258	72.680	30.429	59.560	73.8	56.952

Things to know about time series

values appear in order

so you can't re-arrange the rows!

meaningless without time variable

histograms are not as useful here

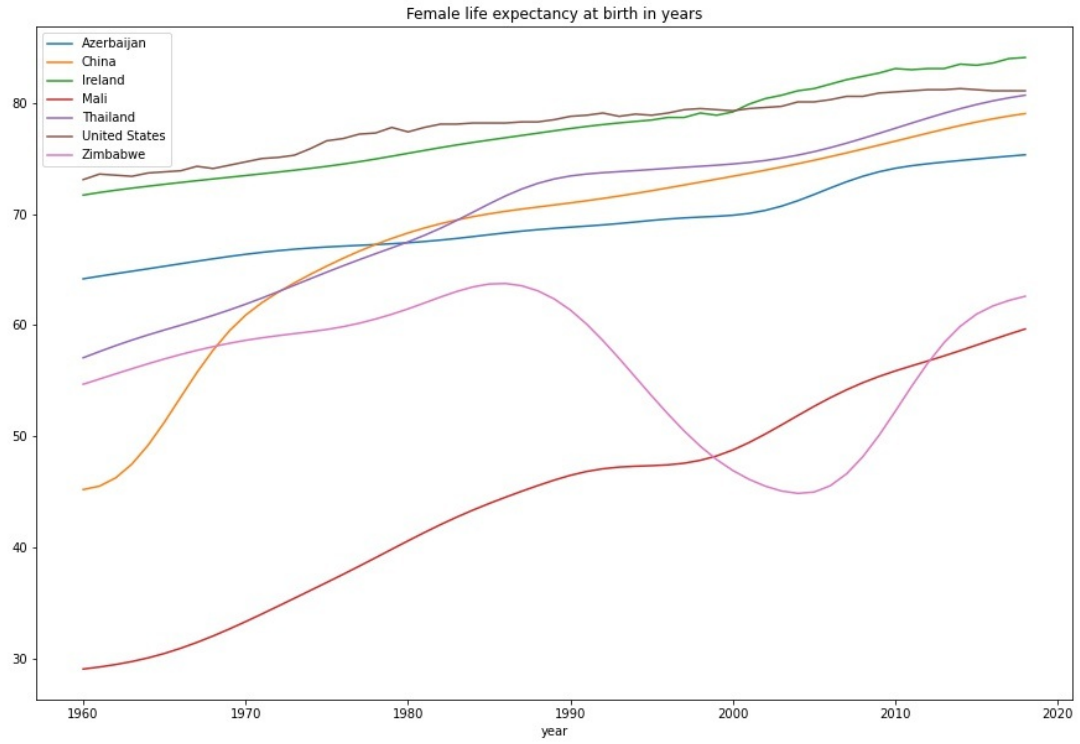
trend is a persistent relationship of variable with time

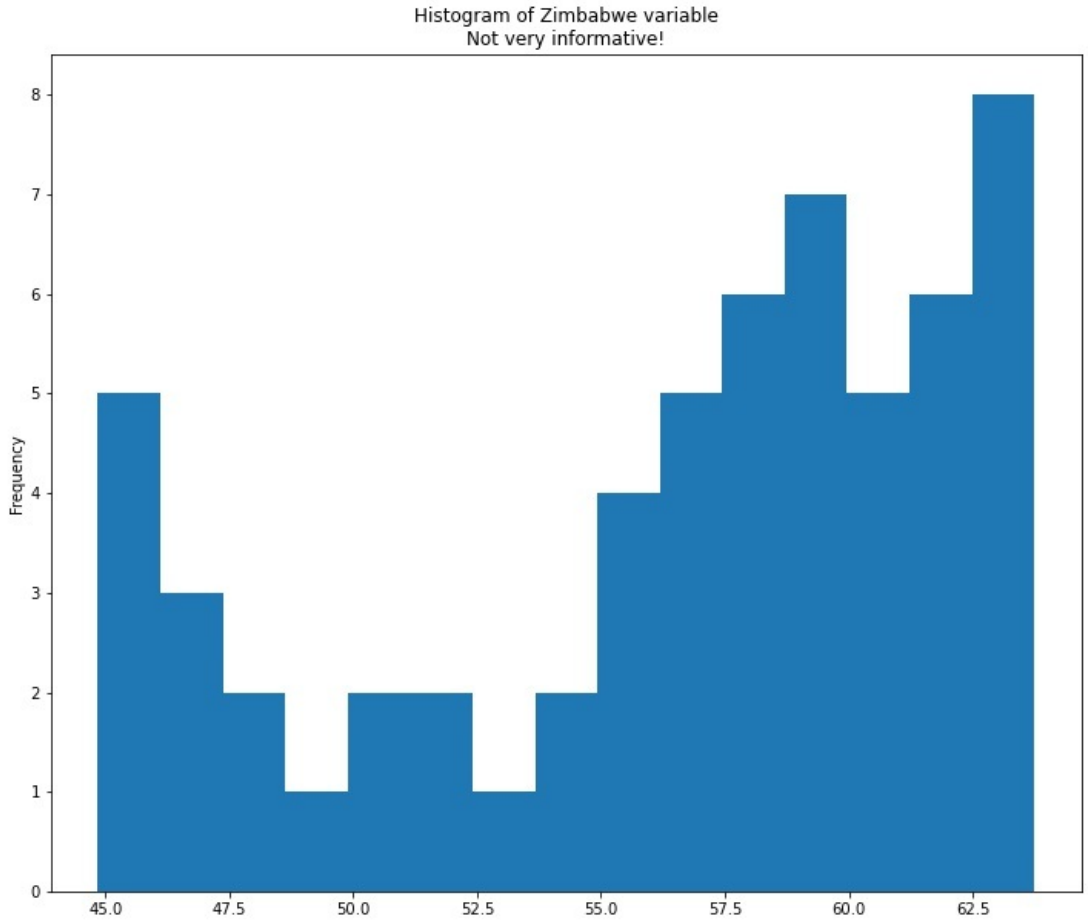
increasing: variable *tends to* increase as time increases

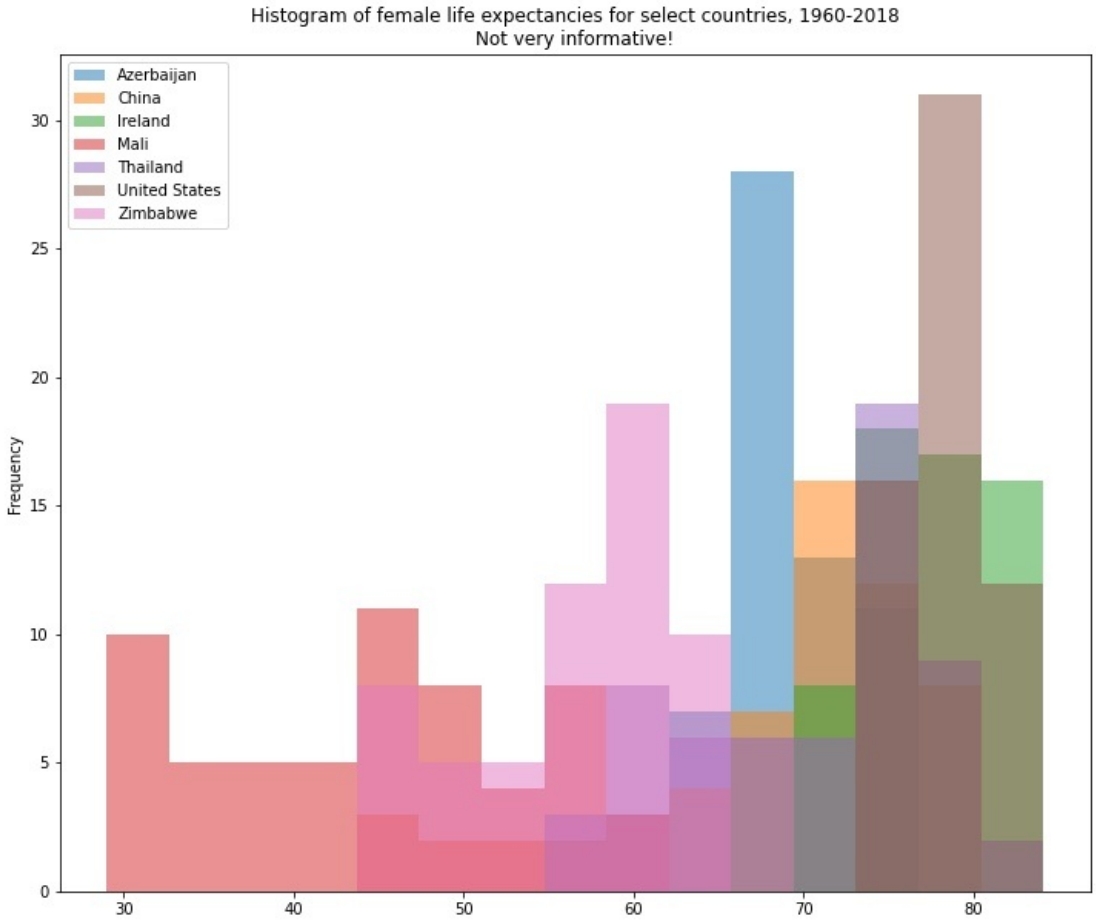
decreasing: variable *tends to* decrease as time increases

This is a time series plot. We'll discuss in the group exercise.

plots created in **notebook on github site**




Histogram of Zimbabwe variable
Not very informative!
Frequency
0
1
2
3
4
5
6
7
8
45.0
47.5
50.0
52.5
55.0
57.5
60.0
62.5


Histogram of female life expectancies for select countries, 1960-2018
Not very informative!
Azerbaijan
China
Ireland
Mali
Thailand
United States
Zimbabwe
Frequency
0
5
10
15
20
25
30
30
40
50
60
70
80

Other ways to view and evaluate time series variables

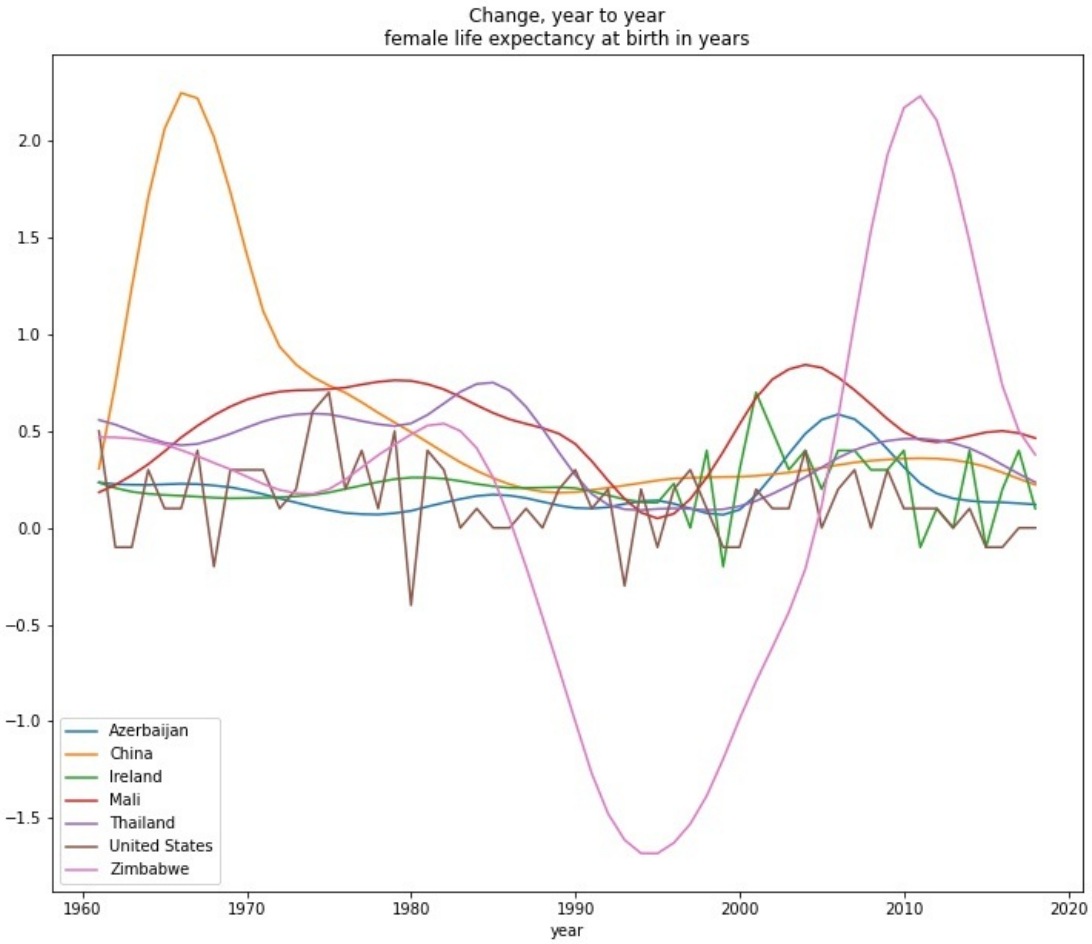
Changes per period, from one time unit to the next, e.g.

life expectancy 1961 − life expectancy 1960

Average change:, i.e. the mean of the changes per perid

average change > 0 can say 'increasing on average', 'increasing trend'

average change < 0 can say 'decr. on average', 'decr. trend'


Change, year to year
female life expectancy at birth in years
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1960
1970
1980
1990
2000
2010
2020
year
Azerbaijan
China
Ireland
Mali
Thailand
United States
Zimbabwe

How to read a time series plots

Trends:

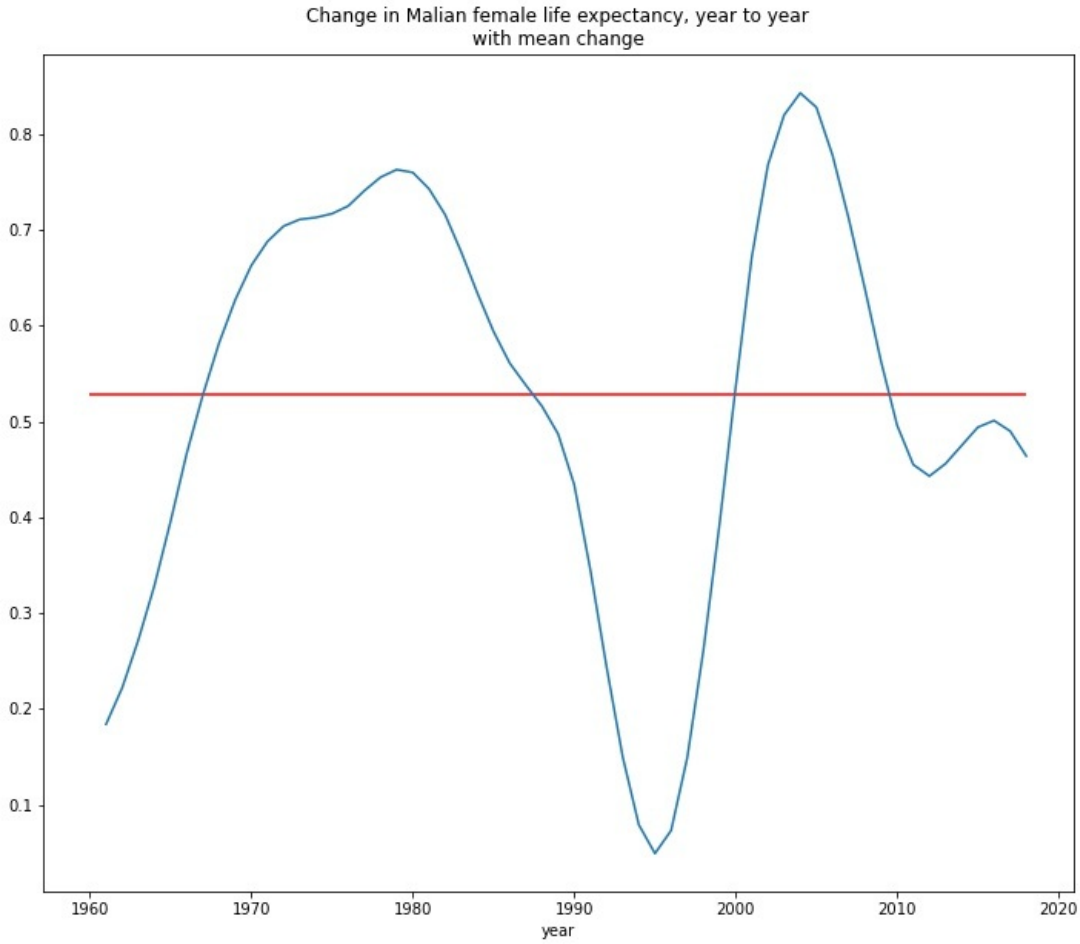
- original time series: look at the **slope** (rise/run)
- positive slope means increasing trend
- ask how well it sticks to a straight line
- **trends over short periods?**

Min, max are easy, quantiles are hard to see

Guess the mean by imagining a *horizontal* line through the middle of the data

Mean of changes per period same as slope of original!

$$\frac{\text{final value} - \text{first value}}{\text{number periods}}$$


Change in Malian female life expectancy, year to year
with mean change
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
1960
1970
1980
1990
2000
2010
2020
year

Scatterplots and correlation

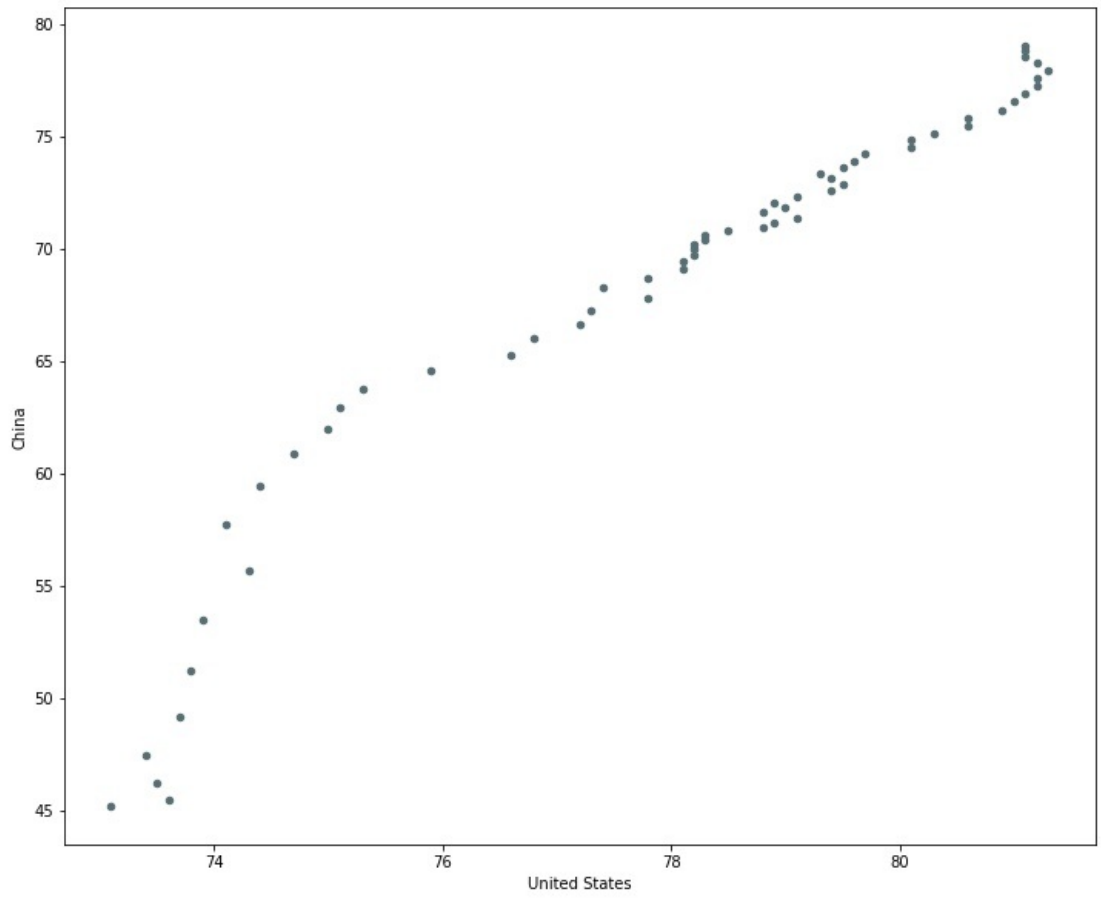
How to evaluate the relationship between any two numeric variables?

Scatterplot

- one variable on the x axis, one on the y axis
- points are pairs of (x, y) values
- describes **how y changes when x changes** and vice versa

Correlation and covariance

- **two-sample statistics** describing **linear relationships**
- meaning, how well does a line through your scatterplot capture its shape?
- more on that later


80
75
70
65
60
55
50
45
China
74
76
78
80
United States

What am I seeing?

x-axis: US life expectancy variable values

y-axis Chinese life expectancy variable values for the same row, meaning same year

United.States	China
73.1	45.191
73.6	45.497
73.5	46.243

How to interpret scatter plots

Do

- "Variable y is (**increasing / decreasing / constant**) in variable x"
- Or, "variable y and x have a (**positive / negative**) relationship"
- "Variables y and x have a (**linear / non-linear**) relationship"

Don't

- "Variable x **causes** variable y to (increase / decrease)"

Strength and type of the relationship

Linear relationships:

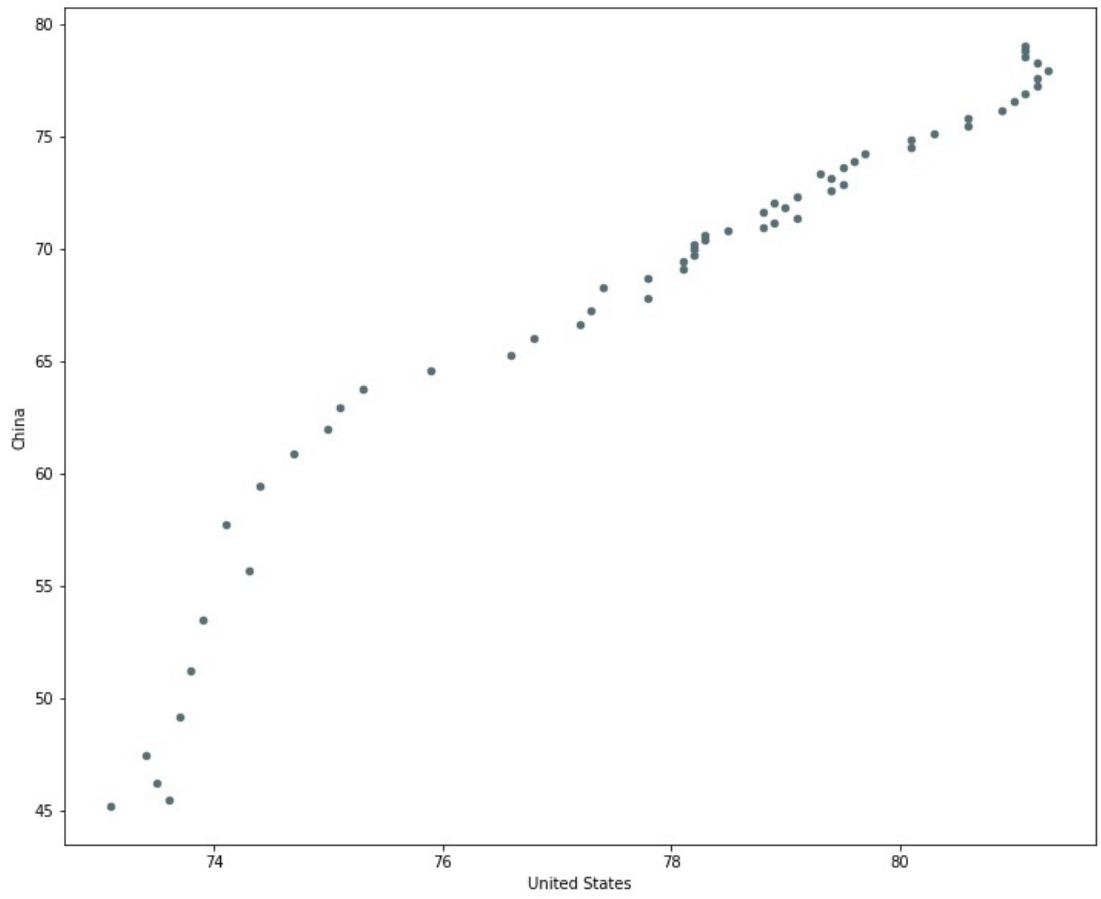
those for which you can put a straight line through the data and

approximately capture its shape.

How straight is straight enough to be 'linear'?

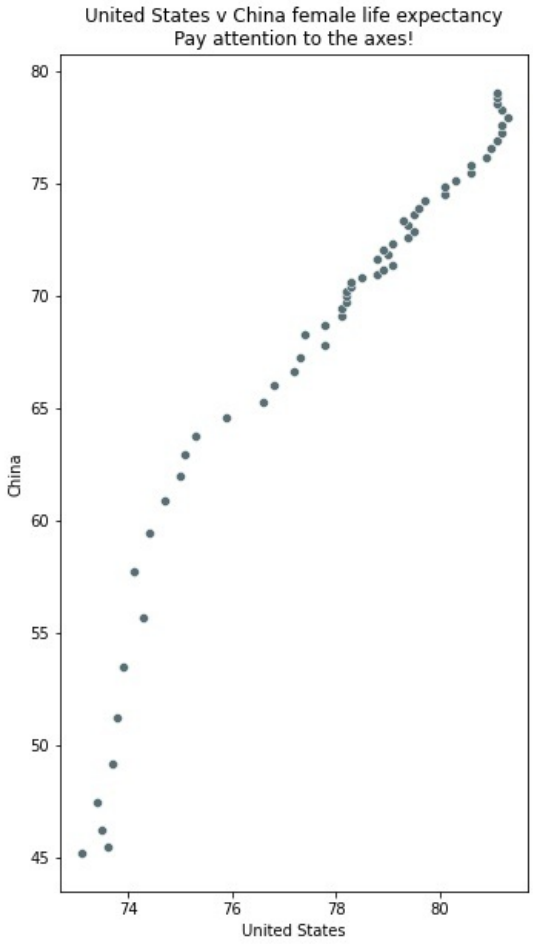
Conclusions based on scatterplots often are open to interpretation!

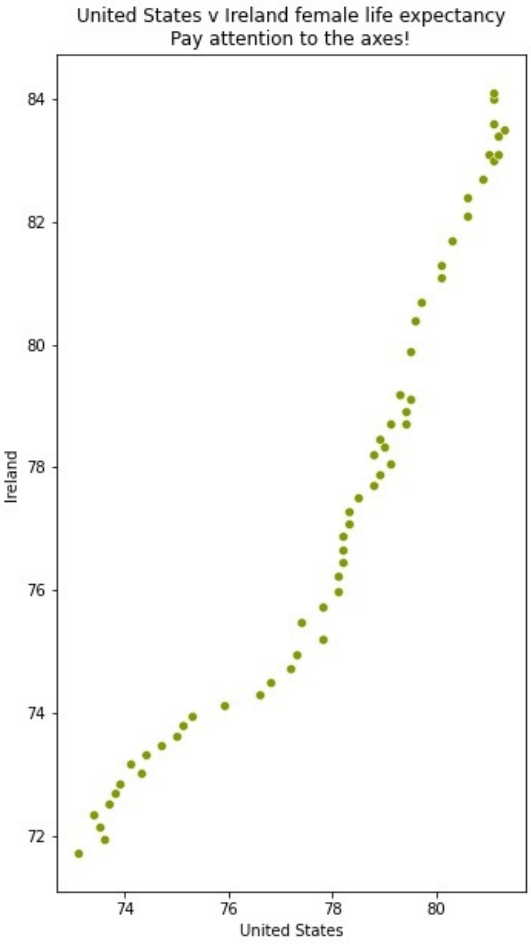
Make this a quantitative evaluation with correlation statistic.

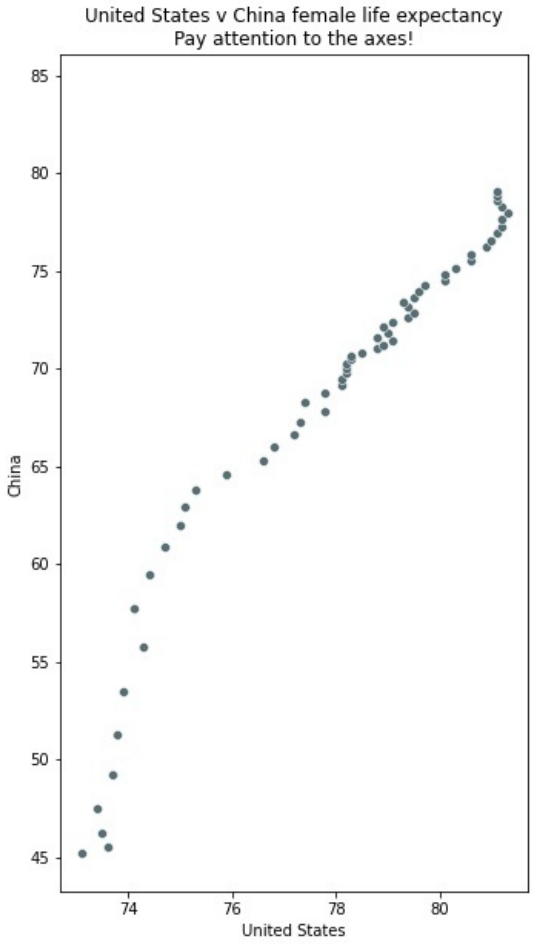

80
75
70
65
60
55
50
45
China
74
76
78
80
United States

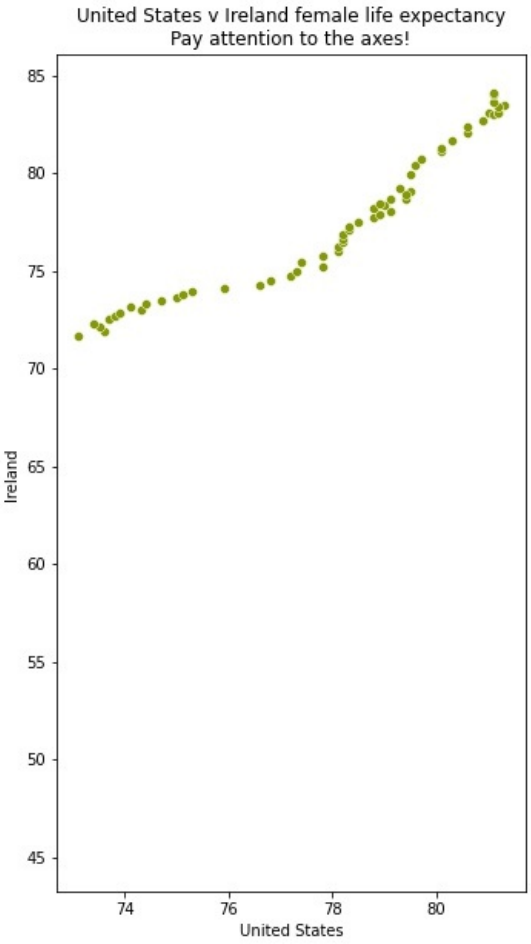
Things you could say

- US and Chinese female life expectancy have a **positive** relationship
- Chinese female life expectancy is **increasing** in US female life expectancy
- US, Chinese female life expectancy have an **somewhat linear** relationship


United States v China female life expectancy
Pay attention to the axes!
China
United States
80
75
70
65
60
55
50
45
74
76
78
80


United States v Ireland female life expectancy
Pay attention to the axes!
Ireland
United States
84
82
80
78
76
74
72
74
76
78
80


United States v China female life expectancy
Pay attention to the axes!
China
United States
85
80
75
70
65
60
55
50
45
74
76
78
80


United States v Ireland female life expectancy
Pay attention to the axes!
Ireland
United States
85
80
75
70
65
60
55
50
45
74
76
78
80

Definitions of some two-sample statistics:

$$X_1, X_2 \ldots X_n, \qquad Y_1, Y_2 \ldots Y_n$$

e.g. $X_1, Y_1 =$ US and Chinese life expectancies in 1960, respectively.

Remember the notation: $\bar{X}$ is the mean and $Var(X)$ is the variance, a measure of how 'spread out' a variable is from its mean.

Standard deviation (one-sample)

The square root of the variance,

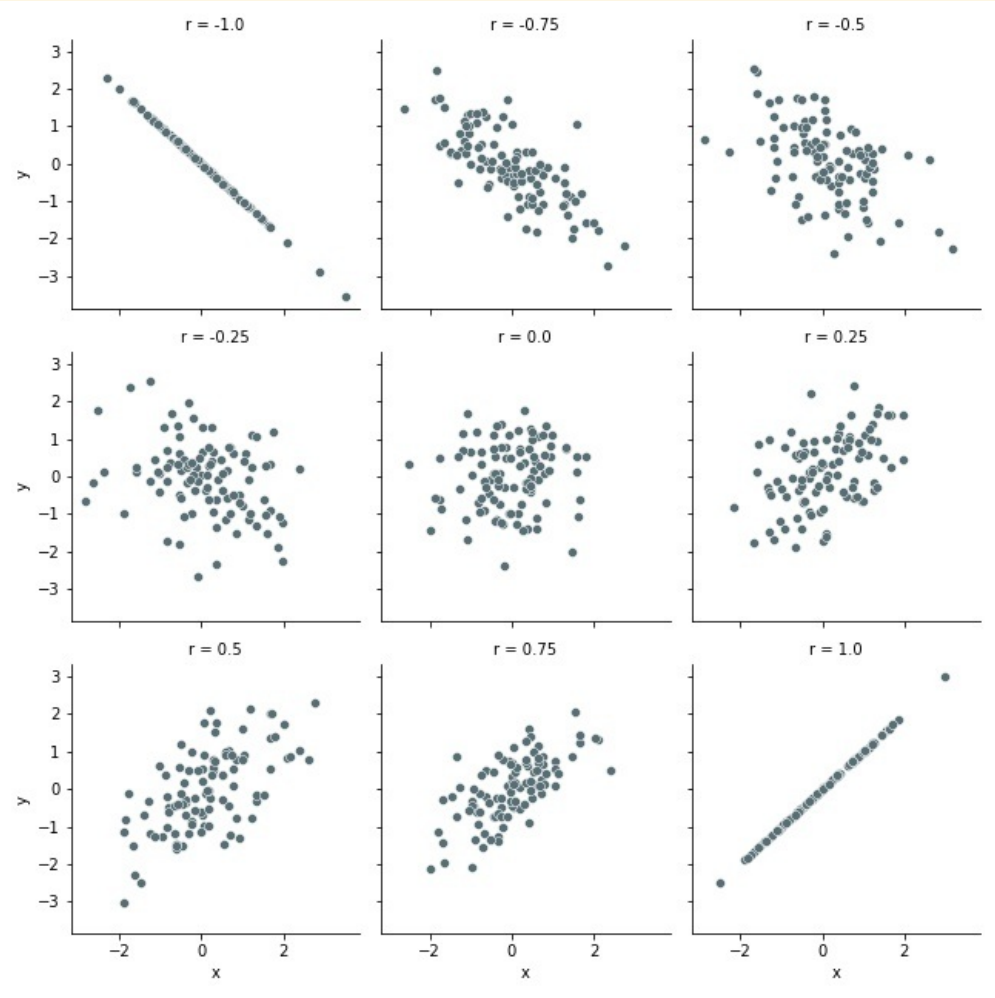
$$\sqrt{Var(X)} = sd(X)$$

Sample covariance

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Sample correlation

$$r = \frac{Cov(X, Y)}{sd(X)sd(Y)}$$


r = -1.0
r = -0.75
r = -0.5
r = -0.25
r = 0.0
r = 0.25
r = 0.5
r = 0.75
r = 1.0
y
x

How to interpret scatterplots in terms of correlation

Strong positive correlation if it looks like $r \approx 1$

almost a straight line with slope 1

Weak correlation if it looks like $r \approx 0$

you cannot easily draw a line to capture the shape

Strong negative correlation if it looks like $r \approx -1$

almost a straight line with slope -1

Group exercise

PollEv.com/brendanbrown849

link to scatterplot for final poll question

wb_lifexpec_scatter_all


Five more minutes