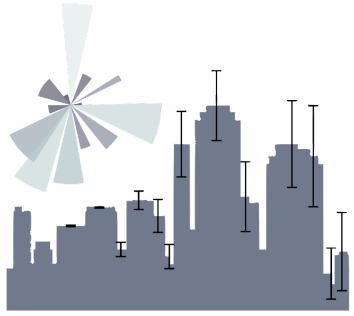
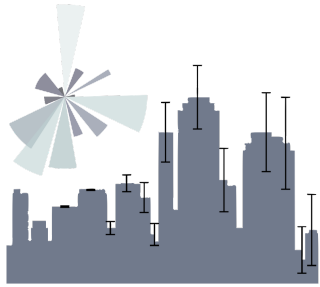
principles of Urban Science 6



residuals, correlation, spatial correlation

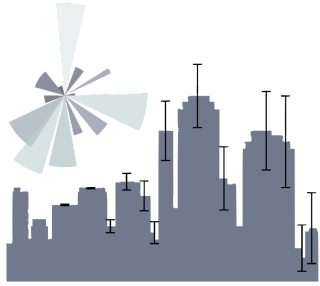
dr.federica bianco | *fbb.space* | *fedhere* | *fedhere*

this slide deck: https://slides.com/federicabianco/pus2020_6



1 residuals

Model residuals

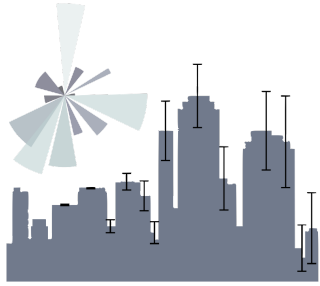


Define Residuals:

model - data.

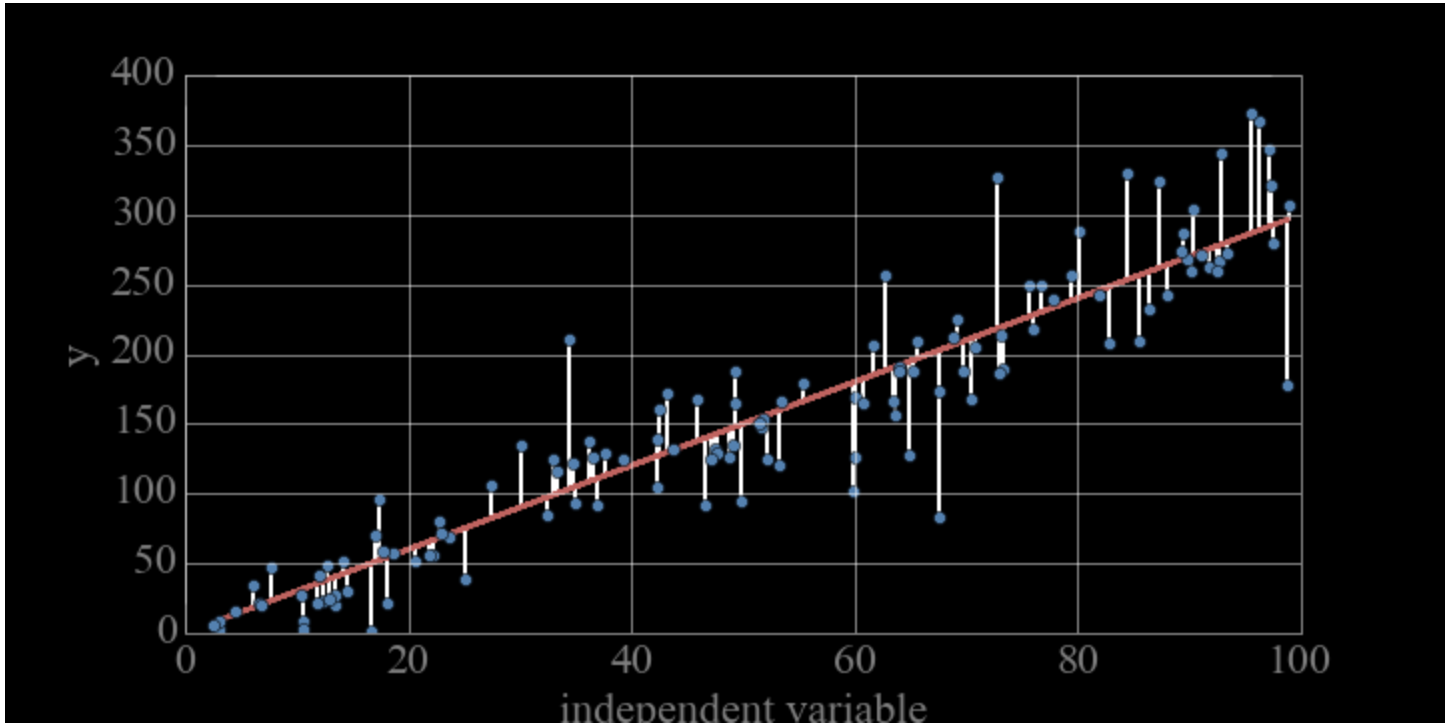
You would like the residuals to be random! that means you captured everything that was not random in the data - everything that you can model.

Model residuals

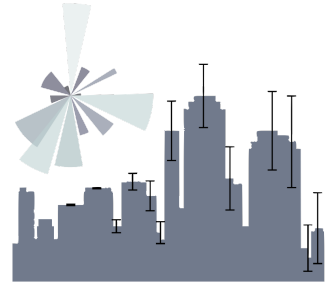


objective function:

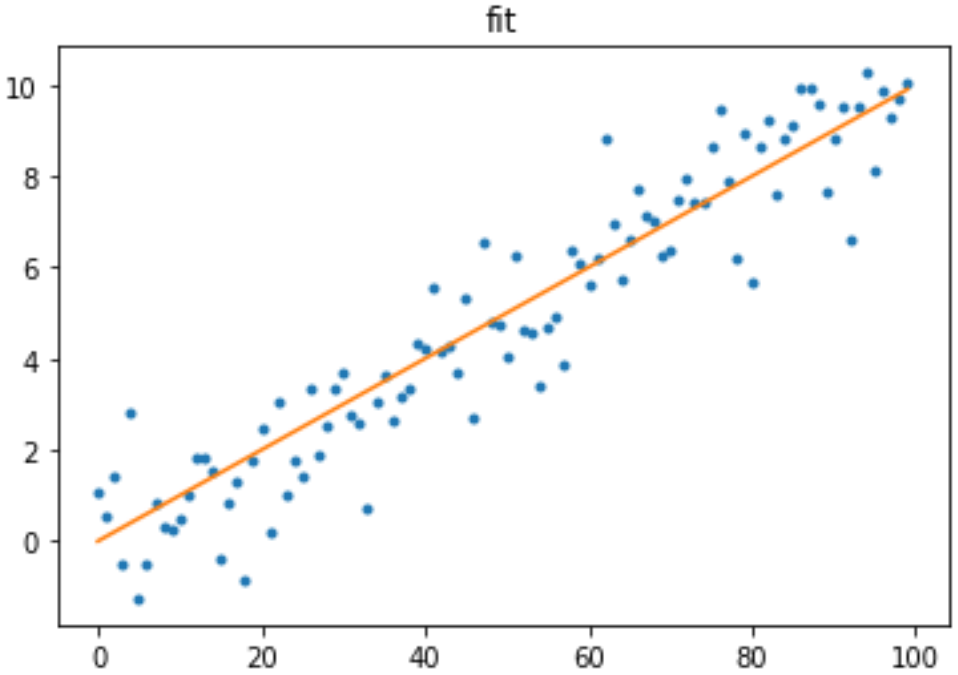
what you want to optimize for

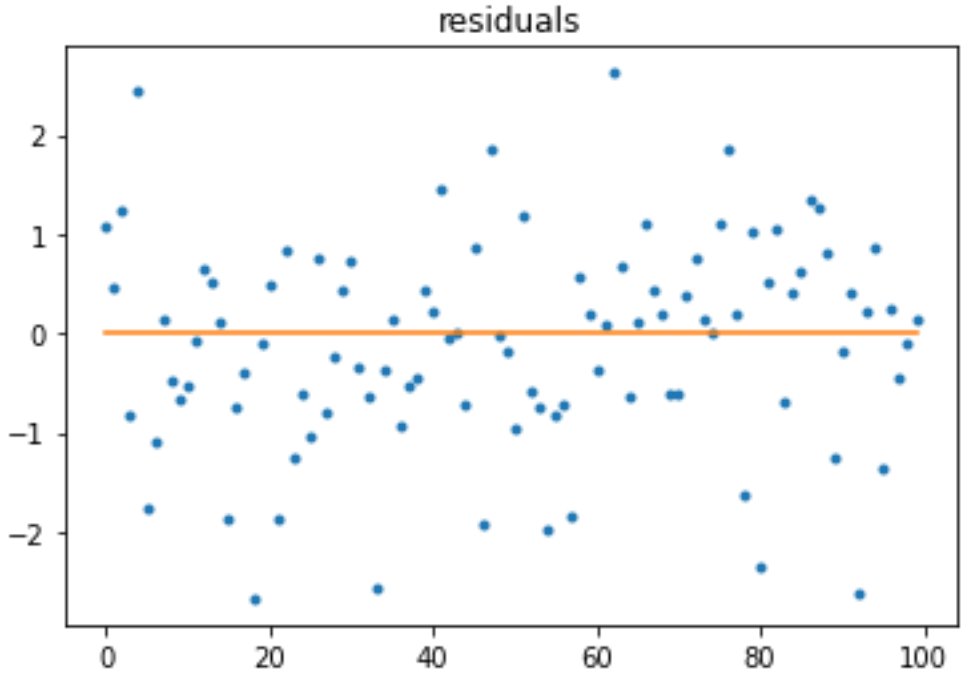


Model residuals

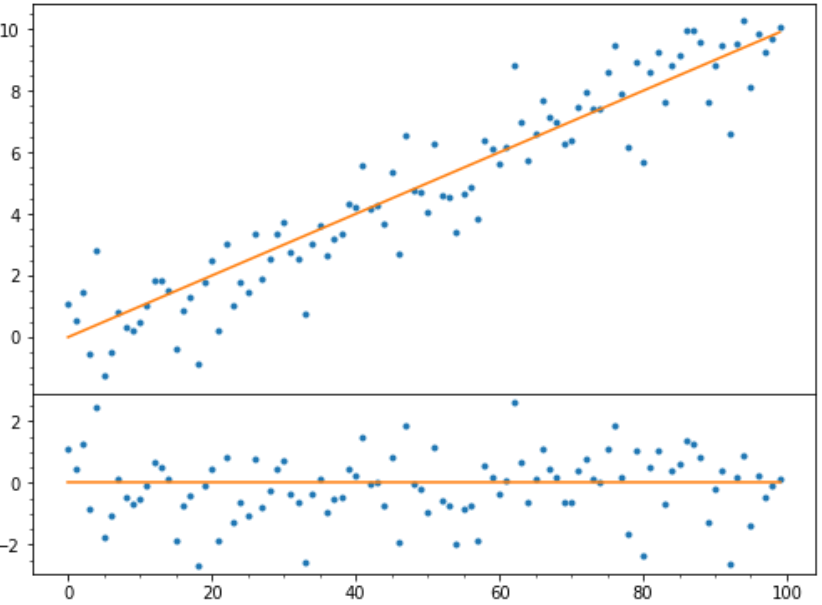


```python
def resplot(x, y, predictions):
  residuals = y - prediction
  return residuals

plt.figure()
plt.plot(x, y, '.')
plt.plot(x, prediction, '-')
plt.title("fit")

plt.figure()
plt.plot(x, resplot(x, y, predictions), '.')
plt.plot(x, np.zeros_like(x), '-')
plt.title("residuals")
```





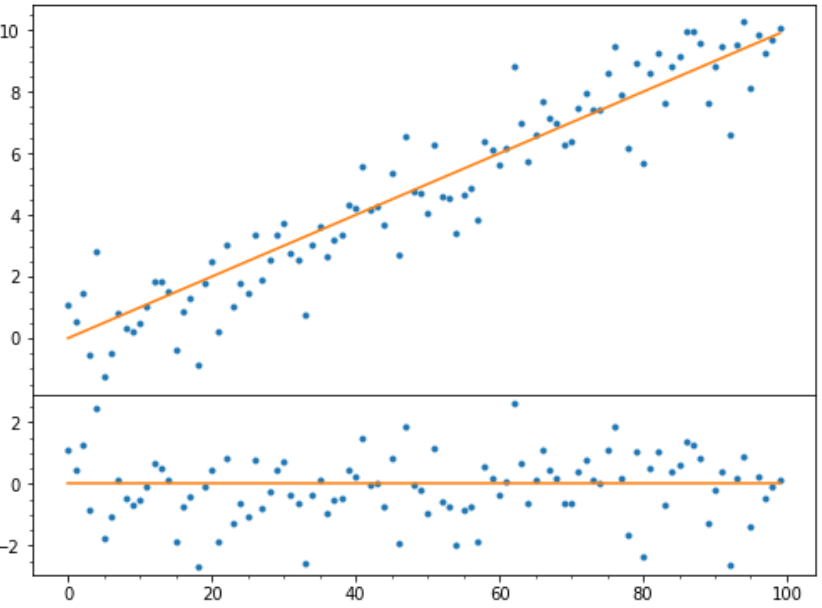
Model residuals

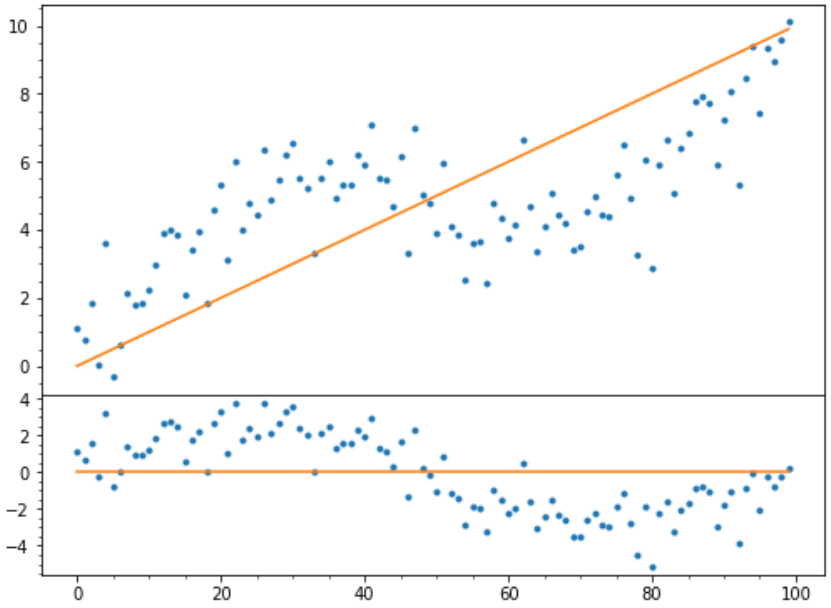


random model residuals

https://github.com/fedhere/PUS2020_FBianco/blob/master/classdemo/residuals_demo.ipynb

Model residuals



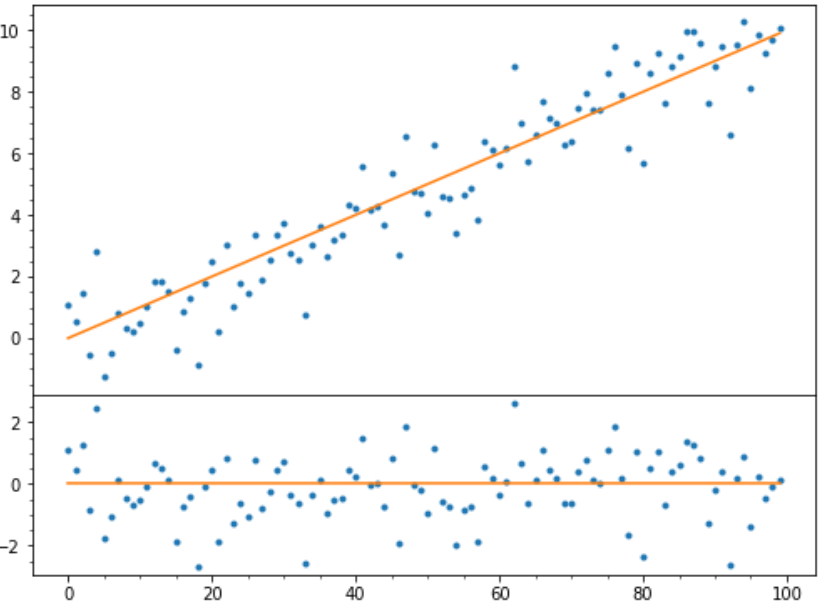


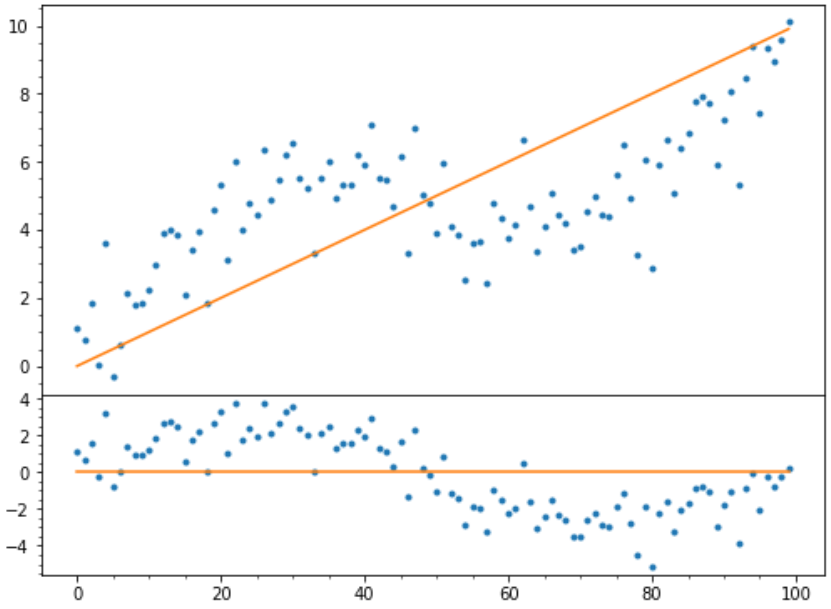
random model residuals

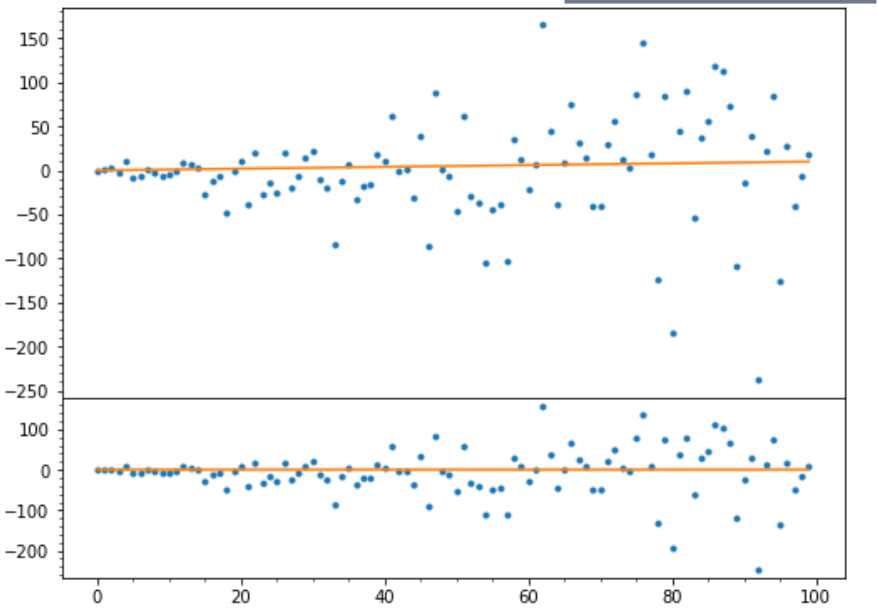
something missed in the model

https://github.com/fedhere/PUS2020_FBianco/blob/master/classdemo/residuals_demo.ipynb

Model residuals







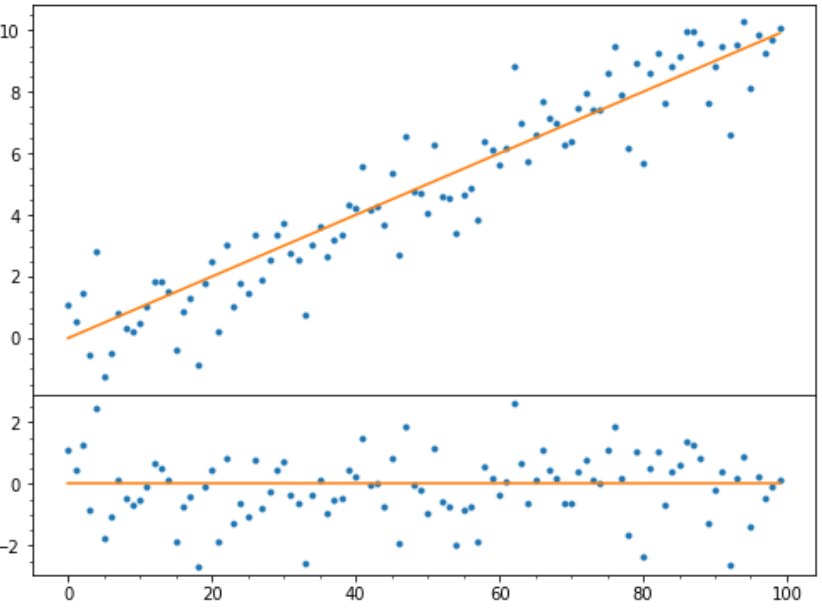
random model residuals

something missed in the model

hetheroscedastic errors

https://github.com/fedhere/PUS2020_FBianco/blob/master/classdemo/residuals_demo.ipynb

Model residuals

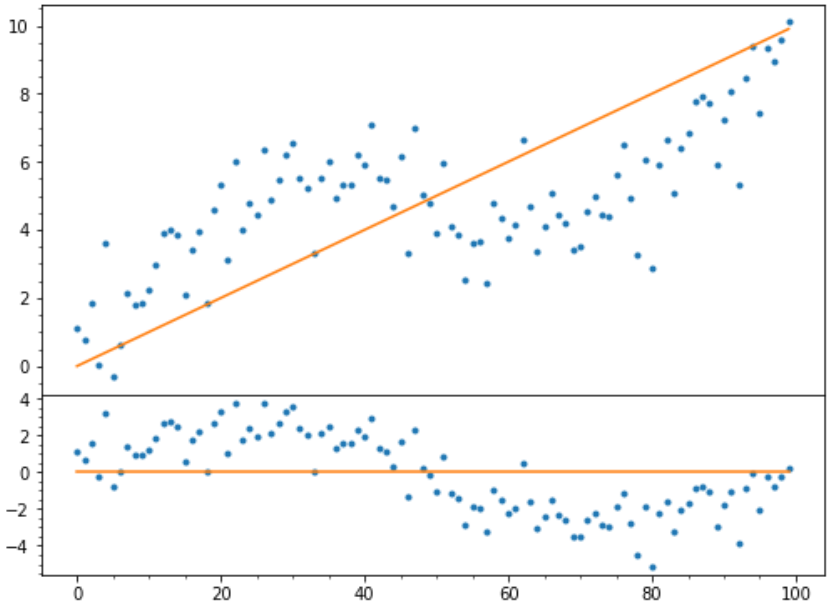


OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.967
Model:	OLS	**Adj. R-squared (uncentered):**	0.967
Method:	Least Squares	**F-statistic:**	2905.
Date:	Mon, 05 Oct 2020	**Prob (F-statistic):**	3.52e-75
Time:	14:38:51	**Log-Likelihood:**	-376.09
No. Observations:	100	**AIC:**	754.2
Df Residuals:	99	**BIC:**	756.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	9.7359	0.181	53.900	0.000	9.377	10.094

Omnibus:	1.594	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.451	**Jarque-Bera (JB):**	1.057
Skew:	0.200	**Prob(JB):**	0.590
Kurtosis:	3.307	**Cond. No.**	1.00

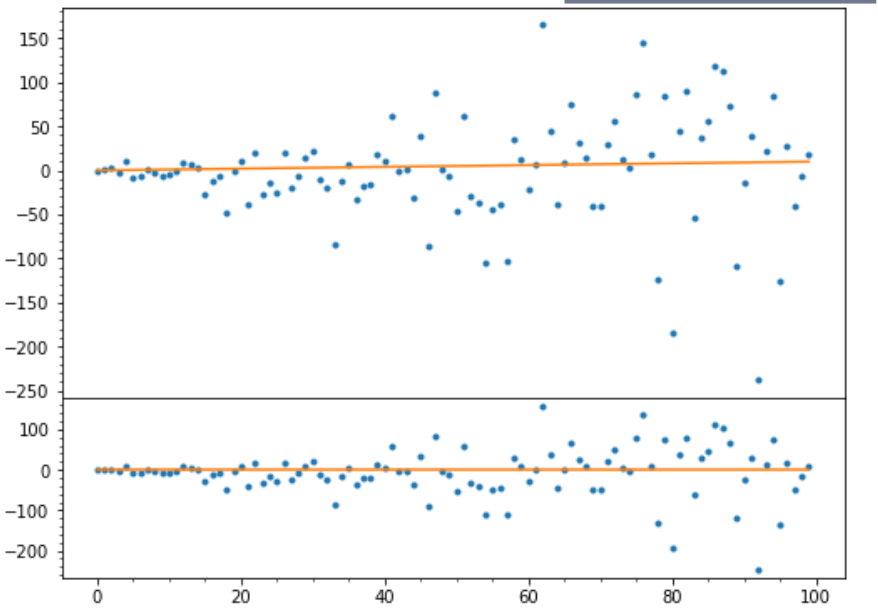


OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.860
Model:	OLS	**Adj. R-squared (uncentered):**	0.859
Method:	Least Squares	**F-statistic:**	608.5
Date:	Mon, 05 Oct 2020	**Prob (F-statistic):**	4.56e-44
Time:	14:41:32	**Log-Likelihood:**	-448.40
No. Observations:	100	**AIC:**	898.8
Df Residuals:	99	**BIC:**	901.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	10.0970	0.409	24.667	0.000	9.285	10.909

Omnibus:	13.460	**Durbin-Watson:**	0.512
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	4.173
Skew:	0.074	**Prob(JB):**	0.124
Kurtosis:	2.010	**Cond. No.**	1.00

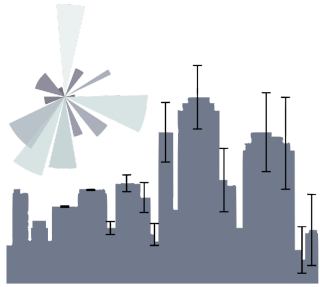


OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.001
Model:	OLS	**Adj. R-squared (uncentered):**	-0.009
Method:	Least Squares	**F-statistic:**	0.08129
Date:	Mon, 05 Oct 2020	**Prob (F-statistic):**	0.776
Time:	14:42:41	**Log-Likelihood:**	-546.69
No. Observations:	100	**AIC:**	1095.
Df Residuals:	99	**BIC:**	1098.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.0277	0.097	0.285	0.776	-0.165	0.220

Omnibus:	28.185	**Durbin-Watson:**	0.002
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5.624
Skew:	0.011	**Prob(JB):**	0.0601
Kurtosis:	1.838	**Cond. No.**	1.00



2 correlation

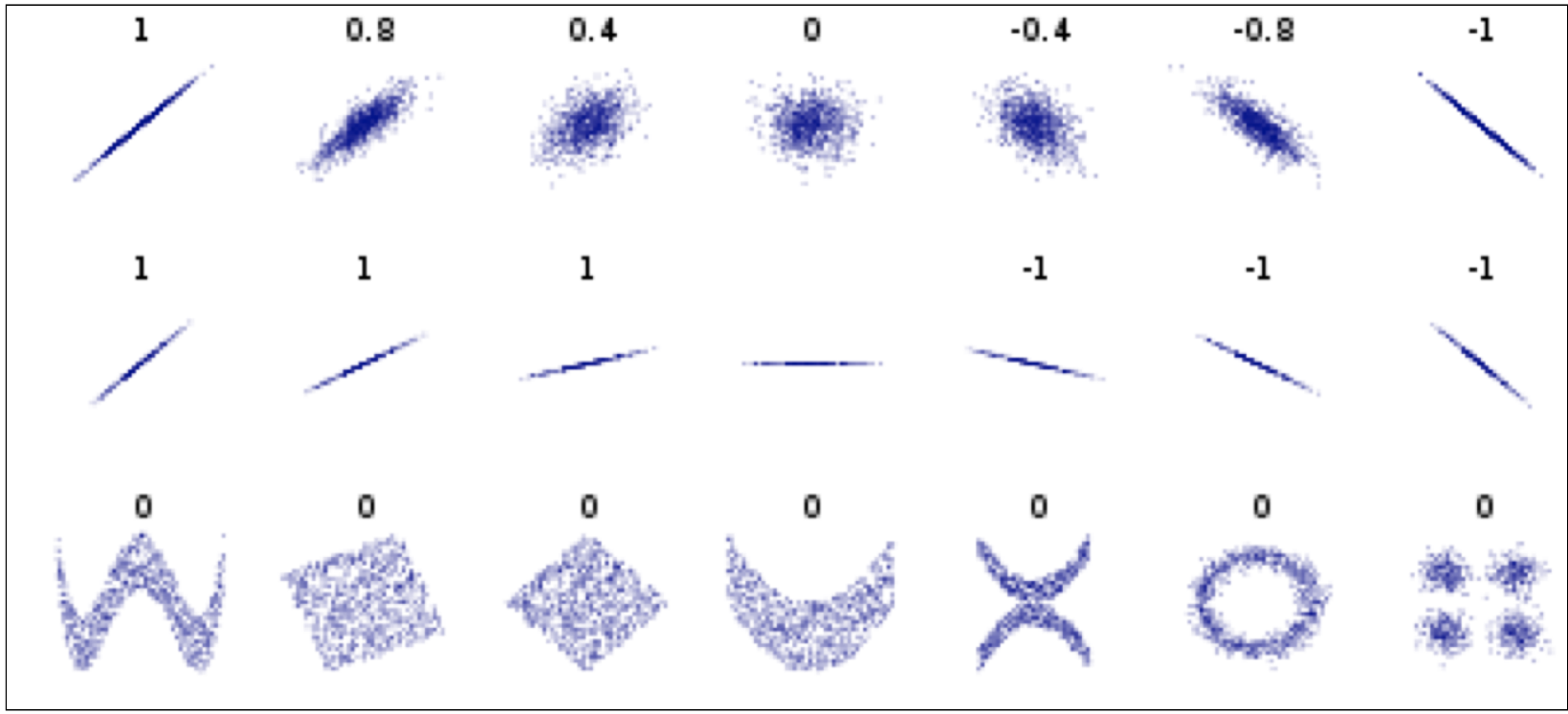
$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Pearson's test: tests *linear* correlation

```
1 import scipy as sp
2 print("Pearson's correlation {:.2}, approximate p-value {:.2}".format(
3     *sp.stats.pearsonr(x, y)))
```

Pearson's correlation 0.94, approximate p-value 2.9e-49



correlation

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Pearson's test

$$\rho_{xy} = 1 - \frac{6 \sum_{i=1}^{N} (x_i - y_i)^2}{n(n^2-1)}$$

Spearman's test

(Pearson's for ranked values)

Choosing the test

Use the table below to choose the test. See below for further details.

How many dichotomous+ (binary) variables?					
0	Both variables interval or ratio?				
	Y	Measures are linear? (No = monotonic*)			
		Y	Pearson correlation		
		N	Spearman correlation		
	N	Both variables are ordinal?			
		Y	Kendall correlation		
		N	Both variables can be ranked?		
			Y	Kendall correlation	
			N	Convert to frequency data and use Chi-square test for independence	
1	Biserial Correlation Coefficient				
2	2 x 2 table? Save the figure				
	Y	Phi			
	N	Cramer's V			

Data has frequency values for each category?	
Y	Chi-square test for independence

+dichotomous = 'can have only two values' (eg. yes/no or 0/1).

*monotonic = constantly increasing or decreasing.

correlation

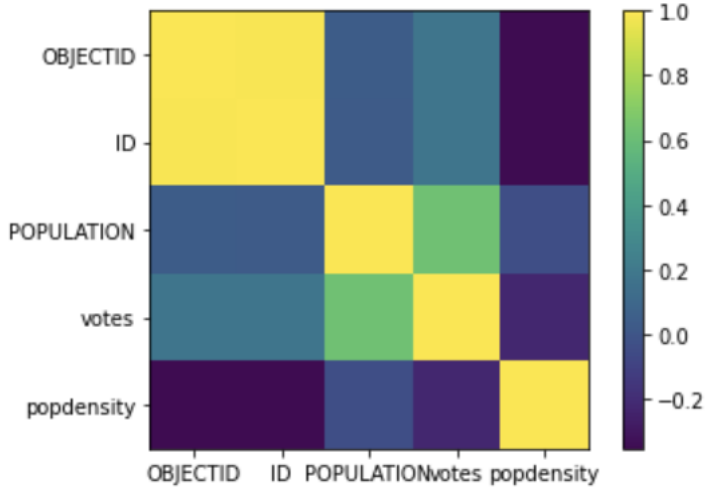
$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \qquad s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Pearson's test

```
de_elecmap.corr()
```

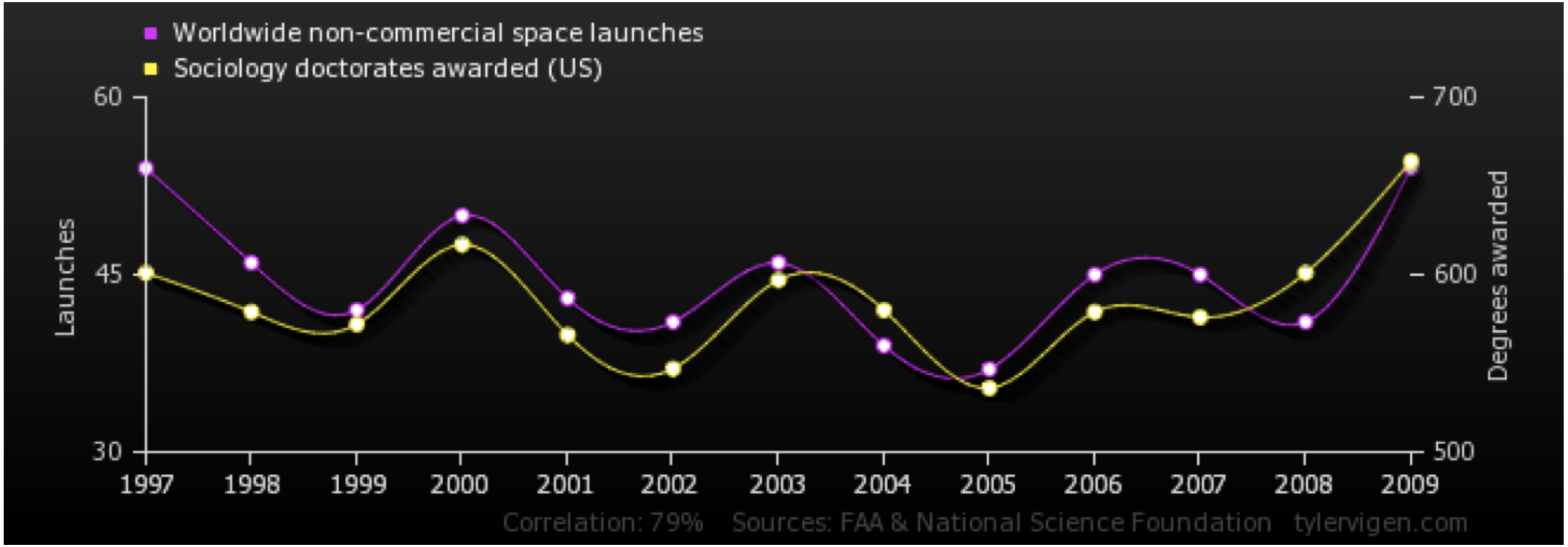
	OBJECTID	ID	POPULATION	votes	popdensity
OBJECTID	1.000000	0.991589	0.037839	0.178269	-0.354172
ID	0.991589	1.000000	0.036542	0.178375	-0.353899
POPULATION	0.037839	0.036542	1.000000	0.624456	-0.032481
votes	0.178269	0.178375	0.624456	1.000000	-0.222647
popdensity	-0.354172	-0.353899	-0.032481	-0.222647	1.000000

```
#Correlation matrix - we will discuss this kind of plot soon
plt.imshow(de_elecmap.corr())
plt.xticks(range(len(de_elecmap.corr())), de_elecmap.corr().columns)
plt.yticks(range(len(de_elecmap.corr())), de_elecmap.corr().columns)
plt.colorbar();
```



correlation

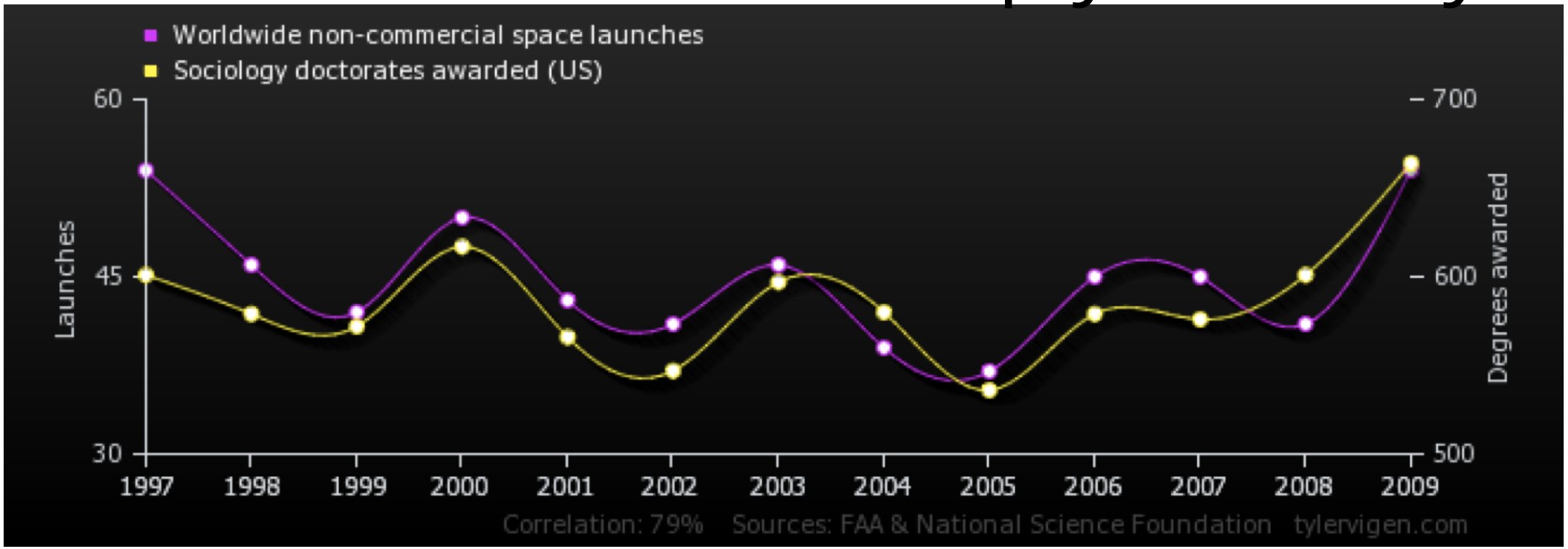
correlation



http://www.tylervigen.com/spurious-correlations

correlation

Correlation does not imply causality



2 things may be related because they share a cause but not cause each other:

icecream sales with temperature

death by drowning with temperature

In the era of big data you may encounter truly spurious correlations

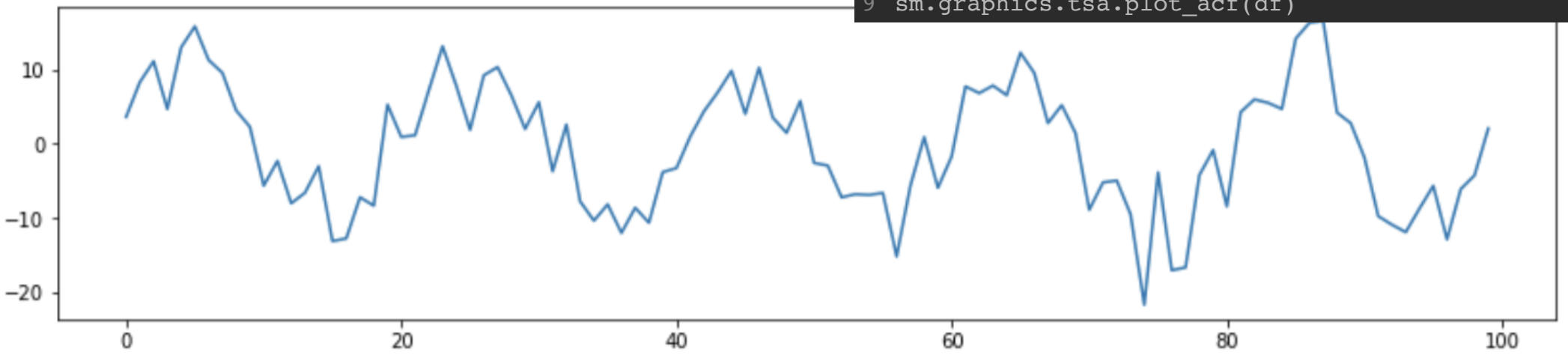
divorce rate in Maine

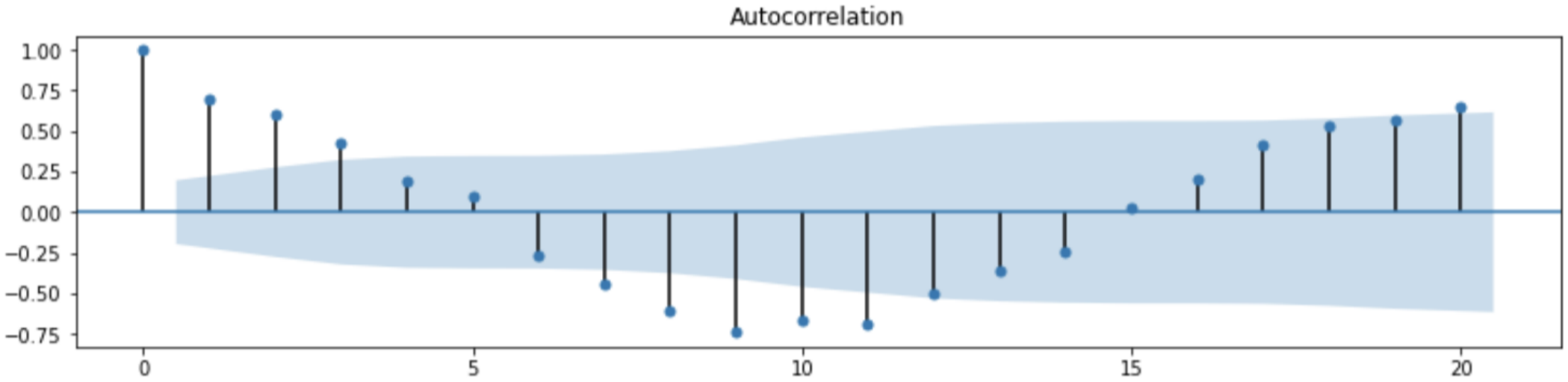
Per capita consumption of Margarine

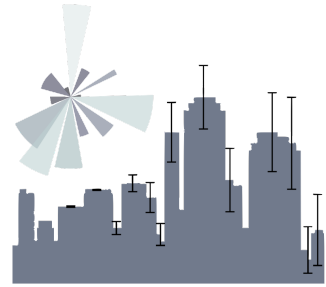
autocorrelation

lag

```python
import statsmodels.api as sm
x = np.random.randn(100) * 3 + np.sin(np.arange(100) / 10 * 3.14)
plt.plot(x)
df = pd.DataFrame({"x":x})
# autocorrelation plots at 5, 10 and 20
df.x.autocorr(lag=10) , df.x.autocorr(lag=20)
     # output (0.0521, -0.791, 0.814)
# make the autocorrelation plot
sm.graphics.tsa.plot_acf(df)
```







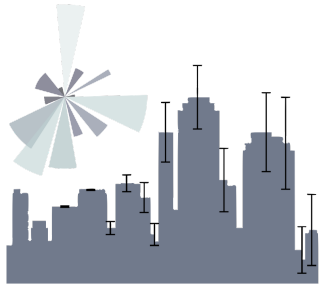
3 spatial correlation

Spatial correlation

Tobler's first law of geography

Everything is related to everything else. But near things are more related than distant things.

This is the first law of Geography introduced by Waldo R. Tobler’s in 1969.



Spatial randomness

The ***Null Hypothesis*** in spatial analysis is commonly that there is no spatial correlation

2 contexts:

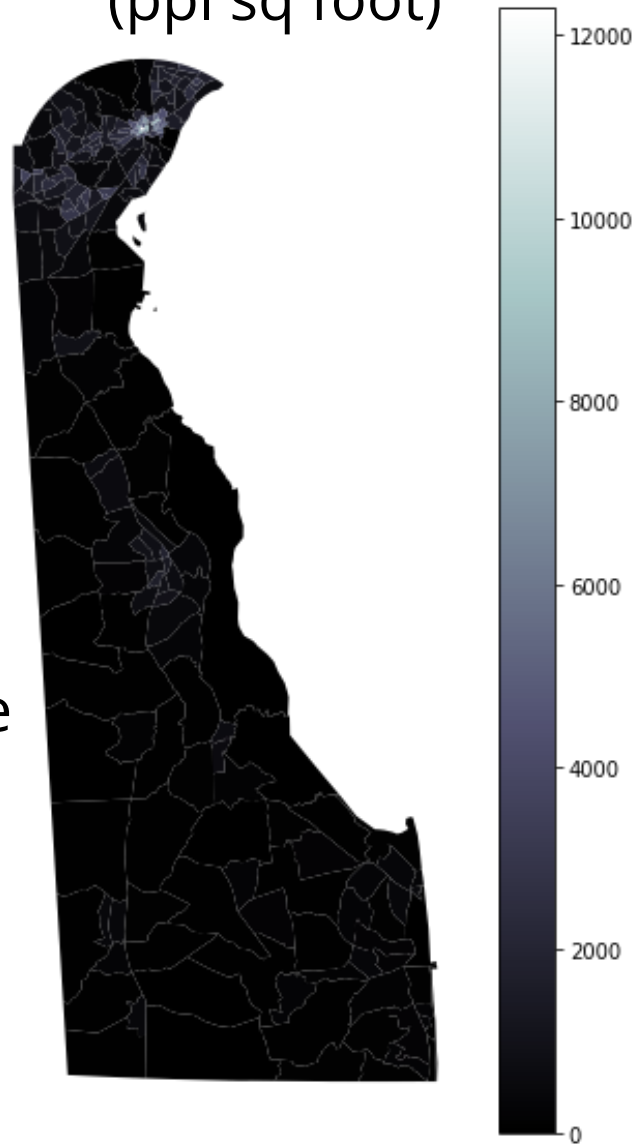
- if you are investigating a phenomenon then randomness is impairing any further analysis
- if you have a model then the residuals to that model should be random

Spatial correlation

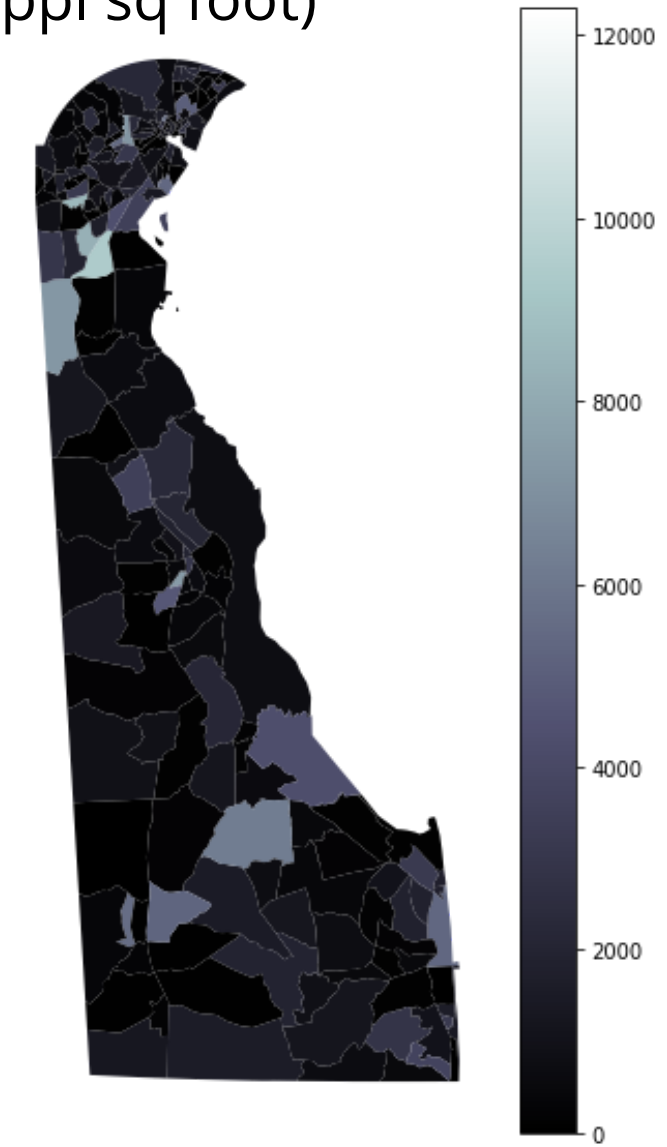
Null Hypothesis: there is no spatial correlation

hard to define! how should this be distributed? depends on size... often we reshuffle variables to compare if the reshuffled version is more random than the non reshuffled

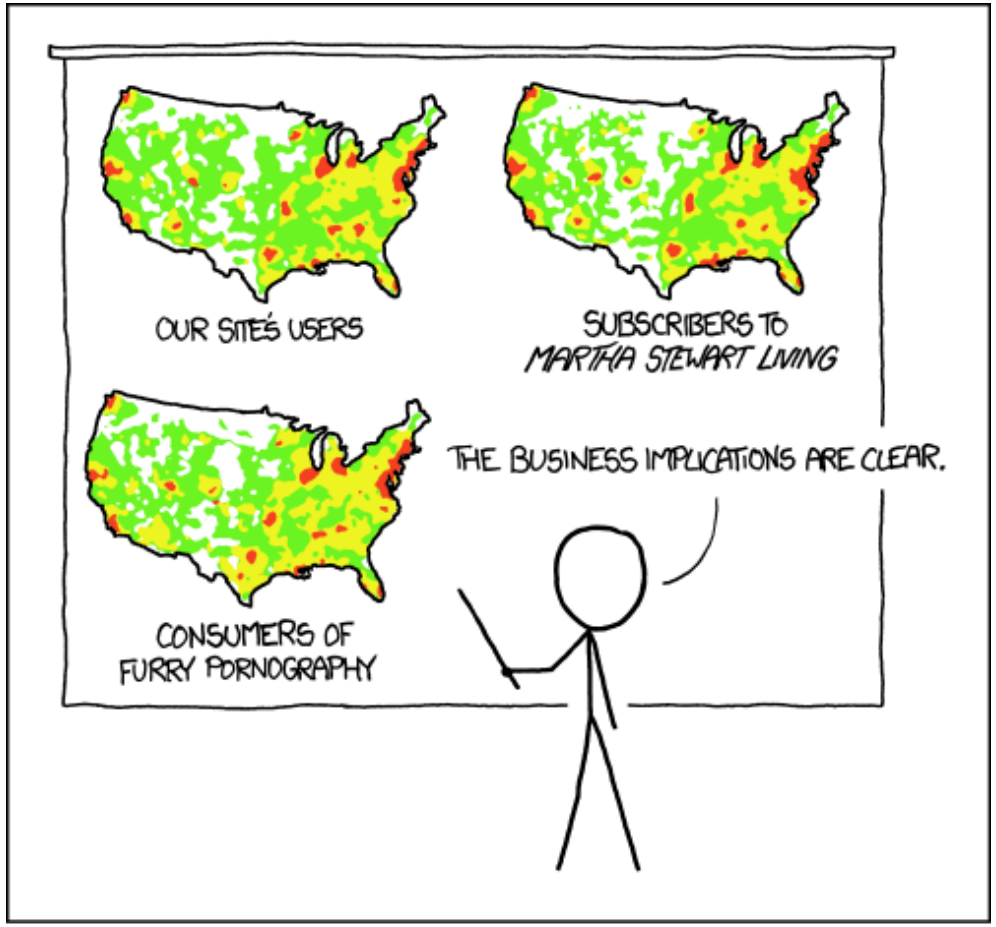
population density DE (ppl sq foot)



reshuffled (random) population density DE (ppl sq foot)



Spatial correlation



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

Spatial index captures spatial structure

attribute similarity b/w 2 variables: f(x, y)

attribute similarity - autocorrelation: f(y_i, y_j)

locational similarity -

does the location influence the correlation

cross product $x \cdot y$

difference $(x - y)^2$ $|x - y|$

$f(y_i, y_j)|\text{location}_i, \text{location}_j$

spatial weights w_{ij}

spatial correlation statistics: $\sum_{i,j} f(y_i, y_j) w_{ij}$

Spatial index captures spatial structure

w_{ij}

- 1 if *i* and *j* are neighbors (e.g.share a border "*contiguity*")
- 0 if *i* and *j* are not neighbors
- 0 if *i* = *j* by convention

$$\begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \cdots & \cdots & \cdots & \\ w_{n,1} & w_{n,2} & \cdots & w_{n,n} \end{bmatrix}$$

spatial correlation statistics: $\sum_{i,j} f(y_i, y_j) w_{ij}$

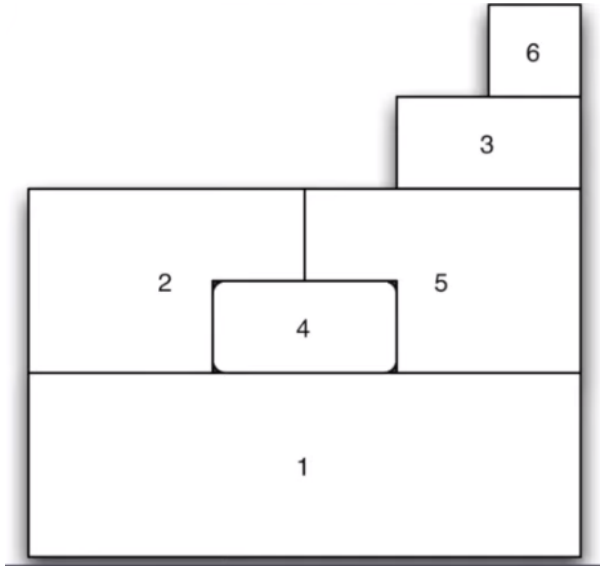
Spatial index captures spatial structure

w_{ij}

1 if *i* and *j* are neighbors
(e.g.share a border "*contiguity*")

0 if *i* and *j* are not neighbors

0 if *i* = *j* by convention



what is the size of the W matrix for this spatial configuration?

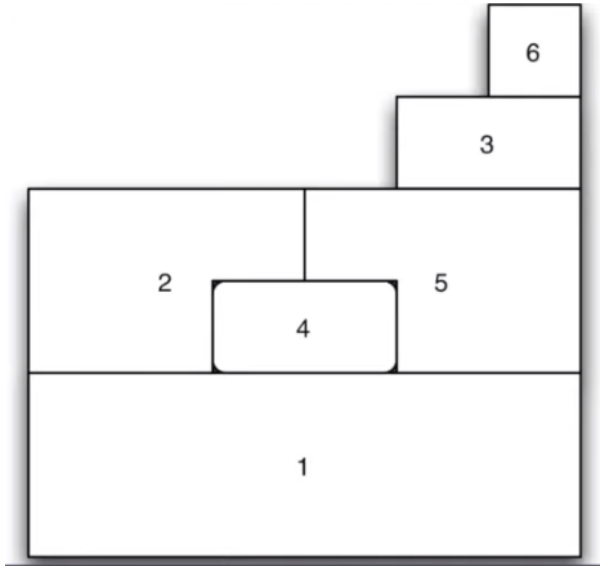
Spatial index captures spatial structure

w_{ij}

1 if *i* and *j* are neighbors
(e.g.share a border "*contiguity*")

0 if *i* and *j* are not neighbors

0 if *i* = *j* by convention

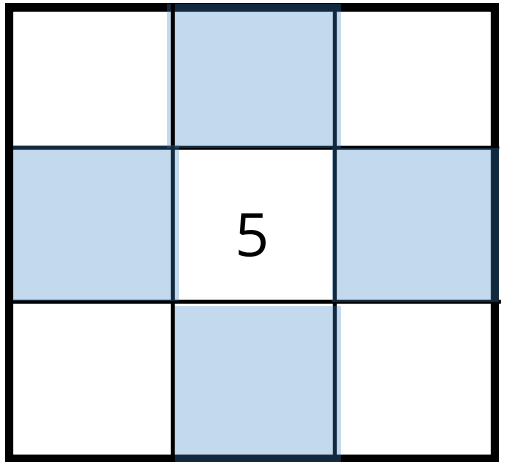


$$\begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Spatial index captures spatial structure

share a boarder

Rook criteria



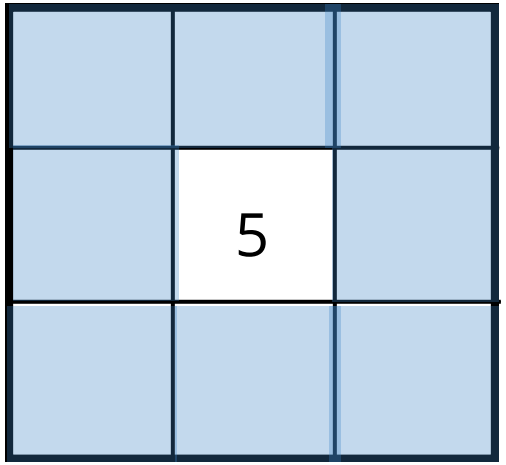
$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Spatial index captures spatial structure

share a boarder

or a corner

Queens criteria



$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Spatial index captures spatial structure

$$I = const. \times d^{-2}$$

can also define based on distance

$$or$$

$$I \propto 1/d^2$$

$$or$$

$$I \propto e^{-d}$$

Distance decay

Distance decay is evident in town/city centres. It can refer to various things which decline with greater distance from the center of the Central Business District (CBD):

- density of pedestrian traffic
- street quality
- quality of shops (depending on definitions of 'quality' and 'center')
- height of buildings
- price of land

Spatial index captures spatial structure

Global autocorrelation

Morans'I

$$I = \frac{\sum_i \sum_j w_{ij} z_i z_i / S_0}{\sum_i z_i^2 / N};$$

$$z_i = y_i - \bar{y}$$

note: the I depends on the weight, so you cannot compare the numerical value of a I

$$z = \frac{I - \bar{I}}{\sigma_I}$$

Standardized I -> z-value can be compared

Moran's I:

<*MI*> = -1/(N-1) => 0 for large numbers

Spatial index captures spatial structure

Global autocorrelation

Morans'I

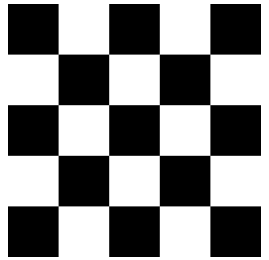
$$I = \frac{\sum_i \sum_j w_{ij} z_i z_i / S_0}{\sum_i z_i^2 / N};$$

$$z_i = y_i - \bar{y}$$

note: the I depends on the weight, so you cannot compare the numerical value of a I

$$z = \frac{I - \bar{I}}{\sigma_I}$$

Standardized I -> z-value can be compared



MI = -1

Moran's I:

<*MI*> = -1/(N-1) => 0 for large numbers

MI = 1

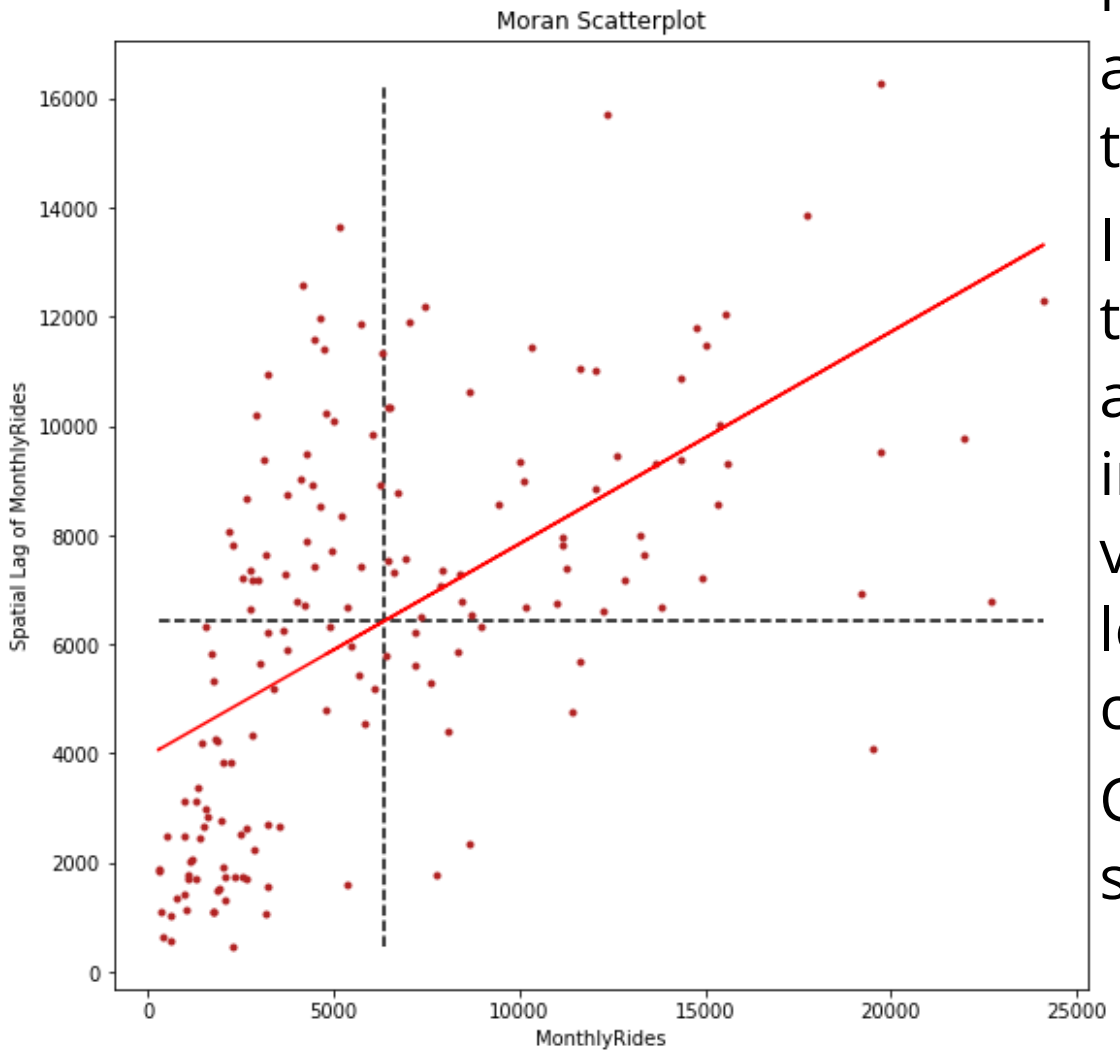
```
print("the global Moran's I is " +
      "{:.2f}\nwhich corresponds to a p-value of the spatial distribution being random {:.2g}".format(
    I_MonthlyRide.I, I_MonthlyRide.p_sim))
```

```
the global Moran's I is 0.39
which corresponds to a p-value of the spatial distribution being random 0.001
```

Spatial index captures spatial structure

Global autocorrelation

Morans'*I*



The Moran scatterplot is an illustration of the relationship between the values of the chosen attribute at each location and the average value of the same attribute at neighboring locations.

In the **upper-right quadrant** are cases where both the value and local average value of the attribute are higher than the overall average value. Similarly, in the lower-left quadrant are cases where both the value and local average value of the attribute are lower than the overall average value. These cases confirm positive spatial autocorrelation.

Cases in the other two quadrants indicate negative spatial autocorrelation.

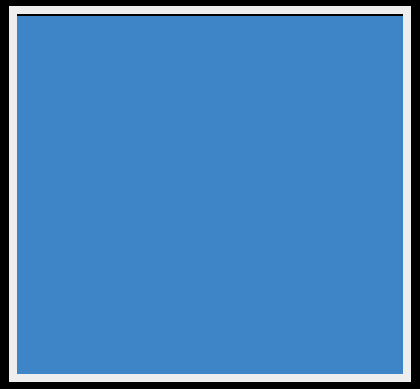
Note: the MI is the slope of the line!

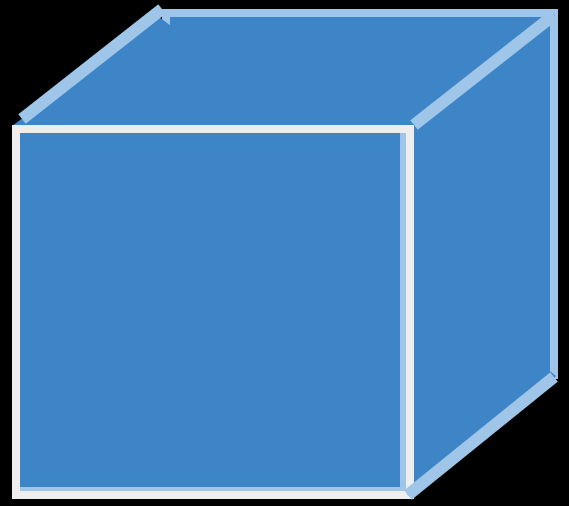
4 scaling laws

quantities that relate by powers

Example:

$L = 1m$





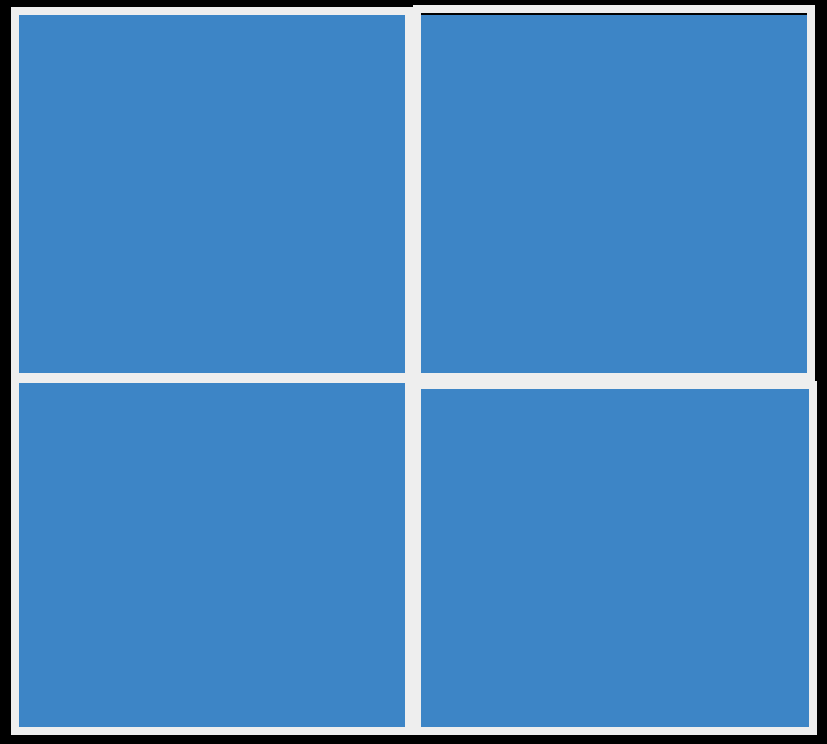
$A = 1m^2$

$V = 1m^3$

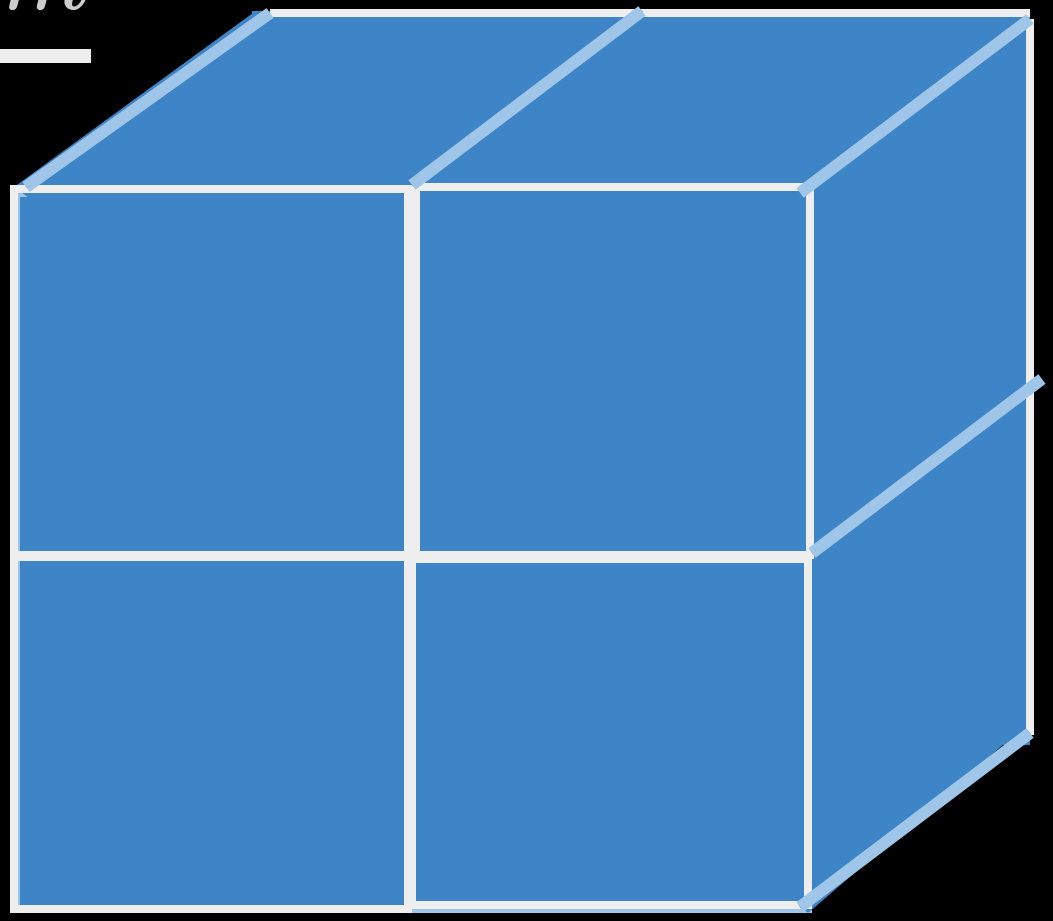
quantities that relate by powers

Example:

$$L = 2\text{x} = 2m$$



$$A = 4\text{x} = 4m^2$$



$$V = 8\text{x} = 8m^3$$

quantities that relate by powers

Example:

scaling law: $(\text{ratio of areas}) = (\text{ratio of lengths})^2$

quantities that relate by powers

Example:

scaling law: $(\text{ratio of areas}) = (\text{ratio of lengths})^2$

scaling law: $(\text{ratio of volumes}) = (\text{ratio of lengths})^3$

quantities that relate by powers

Example:

scaling law: $(\text{ratio of areas}) = (\text{ratio of lengths})^2$

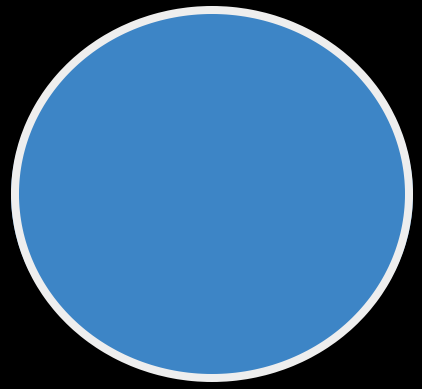
scaling law: $(\text{ratio of volumes}) = (\text{ratio of lengths})^3$

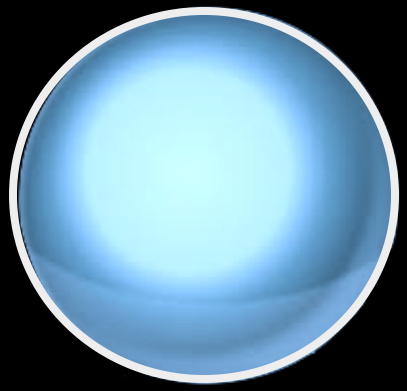
regardless of the shape!

quantities that relate by powers

Example:

$r = 1m$





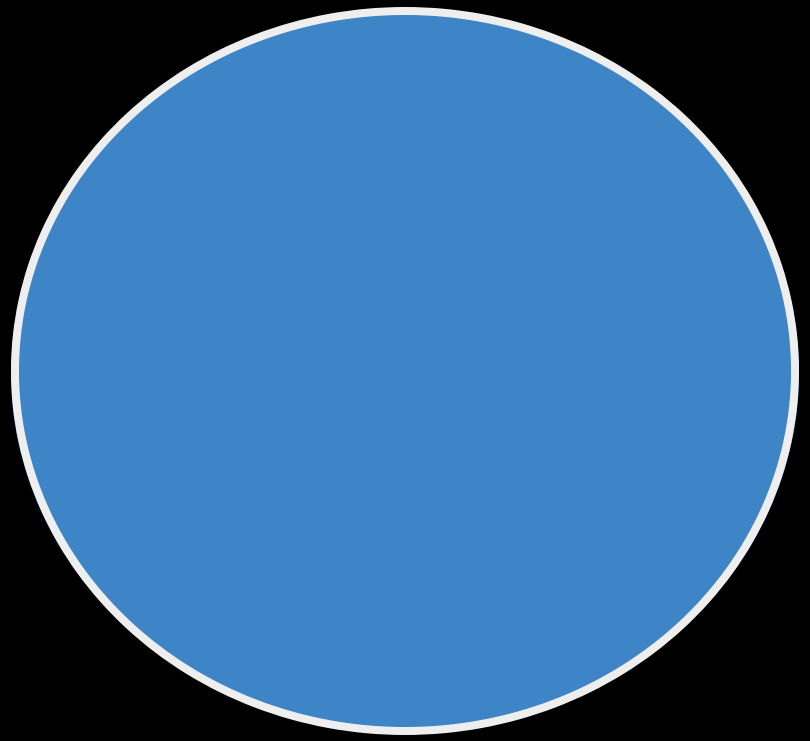
$A = 1m^2$

$V = 1m^3$

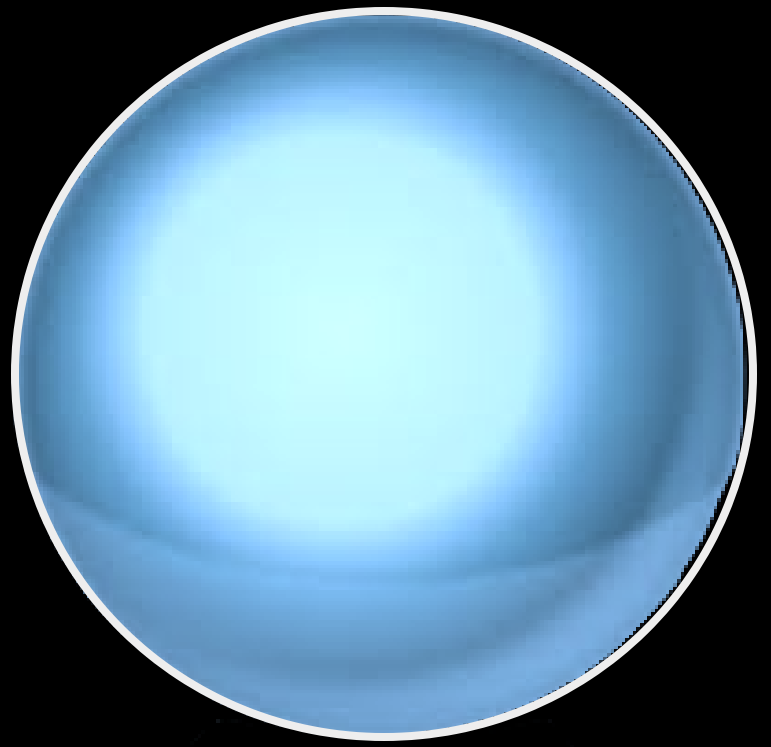
quantities that relate by powers

Example:

$r = 1m$



$V \sim 4\text{x}, V = \text{const}\ r^2$



$V \sim 8\text{x}, V = \text{const}\ r^3$

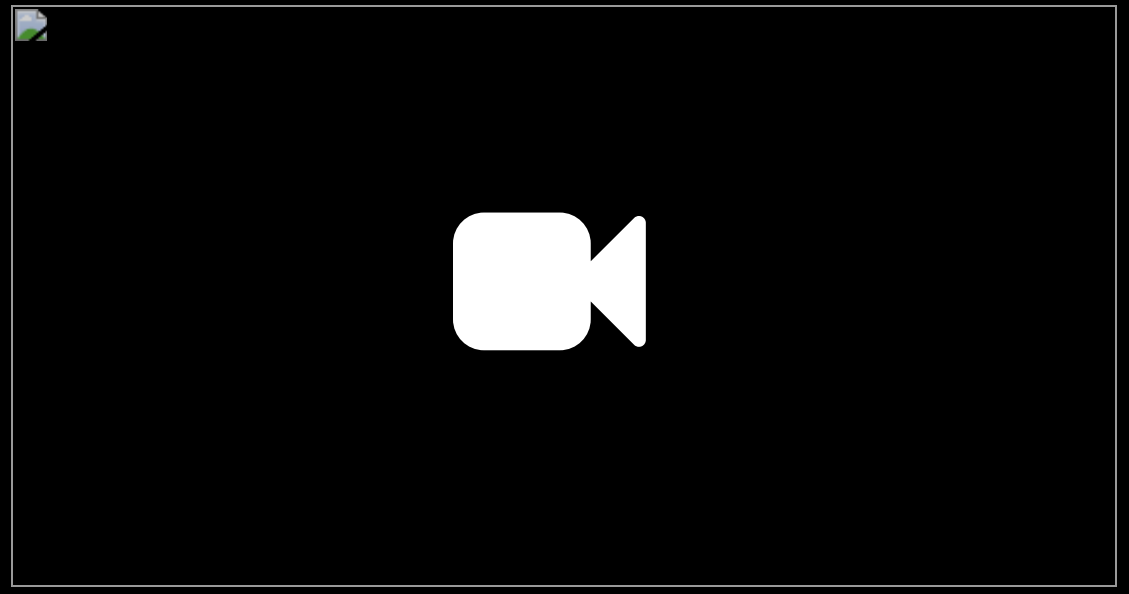
why is this important?

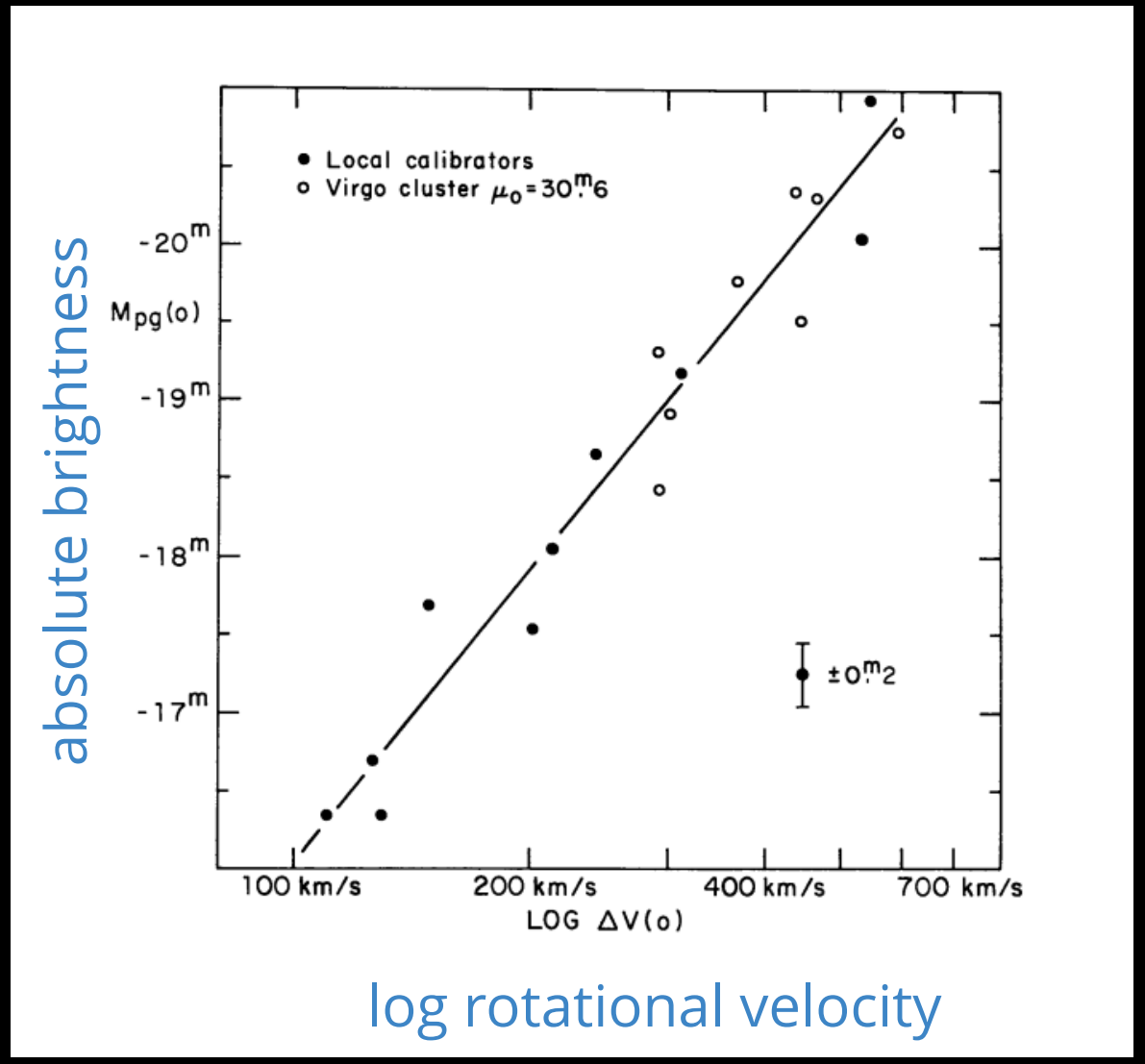
The exsitence of a **scaling** relationship between quantities reveals an underlying driving mechanism

Astrophysics

The **Tully–Fisher relation** is an *empirical relationship between the intrinsic luminosity of a spiral galaxy and its torational velocity*

R. Brent **Tully** and J. Richard **Fisher, 1977**
Astronomy and Astrophysics, 54, 661



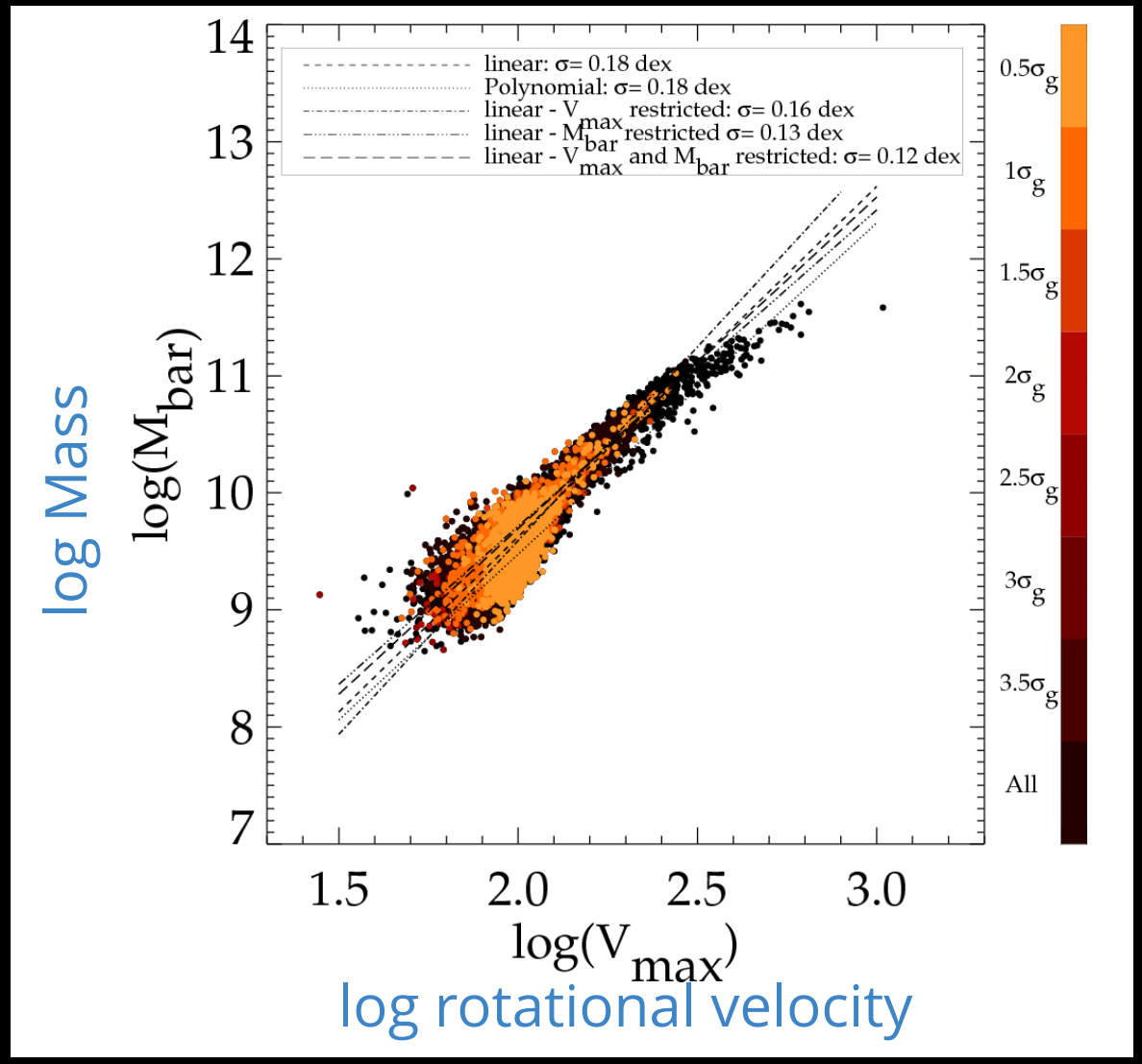


Astrophysics

The **Tully–Fisher relation** is an *empirical relationship between the intrinsic luminosity of a spiral galaxy and its torational velocity*

R. Brent **Tully** and J. Richard **Fisher, 1977**

GRAVITY



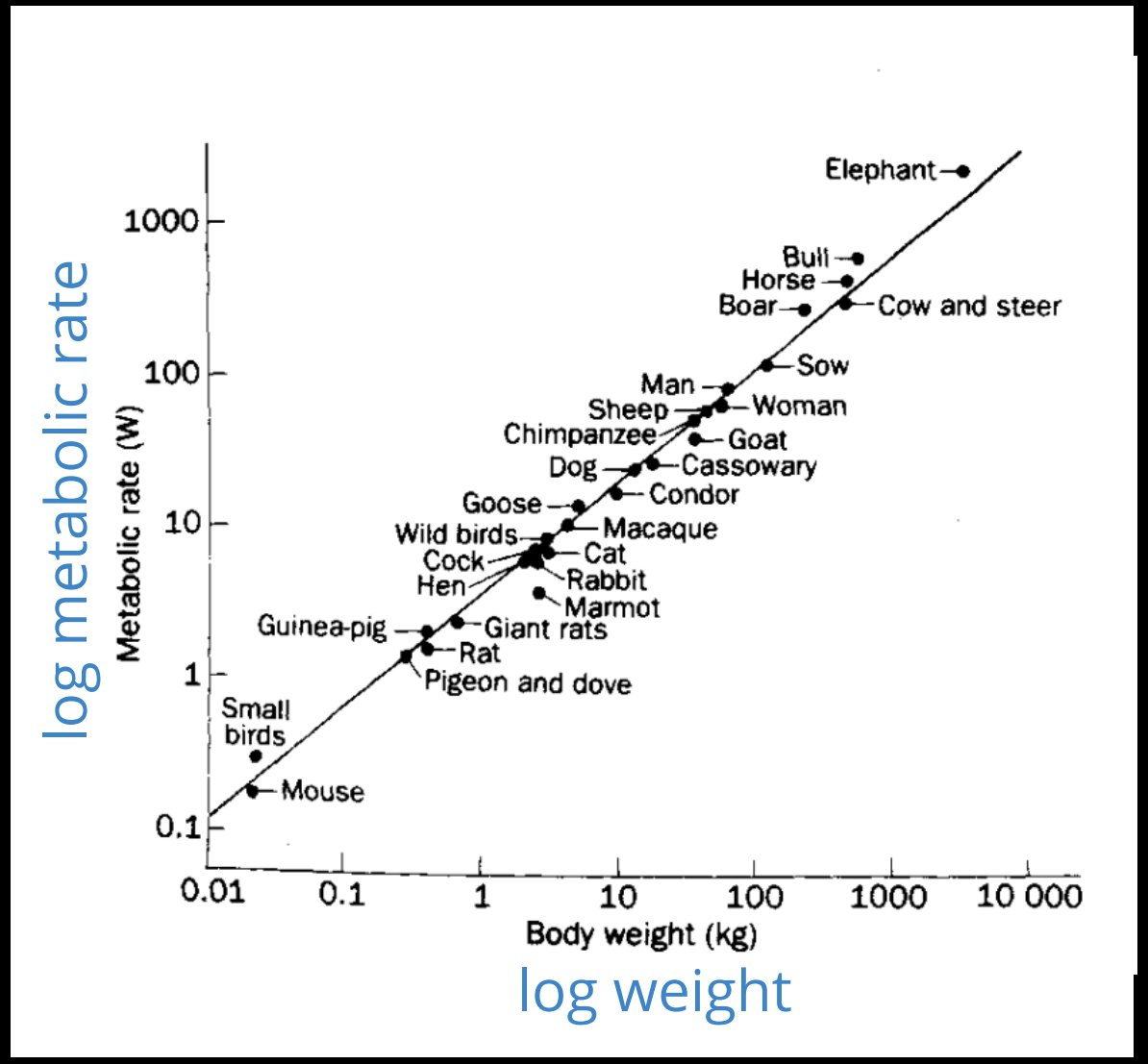
Sorce Jenny *et al.*

Biology

Basal metabolism of mammals (that is, the minimum rate of energy generation of an organism) has long been known to scale empirically as

$$B \propto M^{3/4}$$

KLEIBER, M. (1932). Body size and metabolism. Hilgardia 6, 315



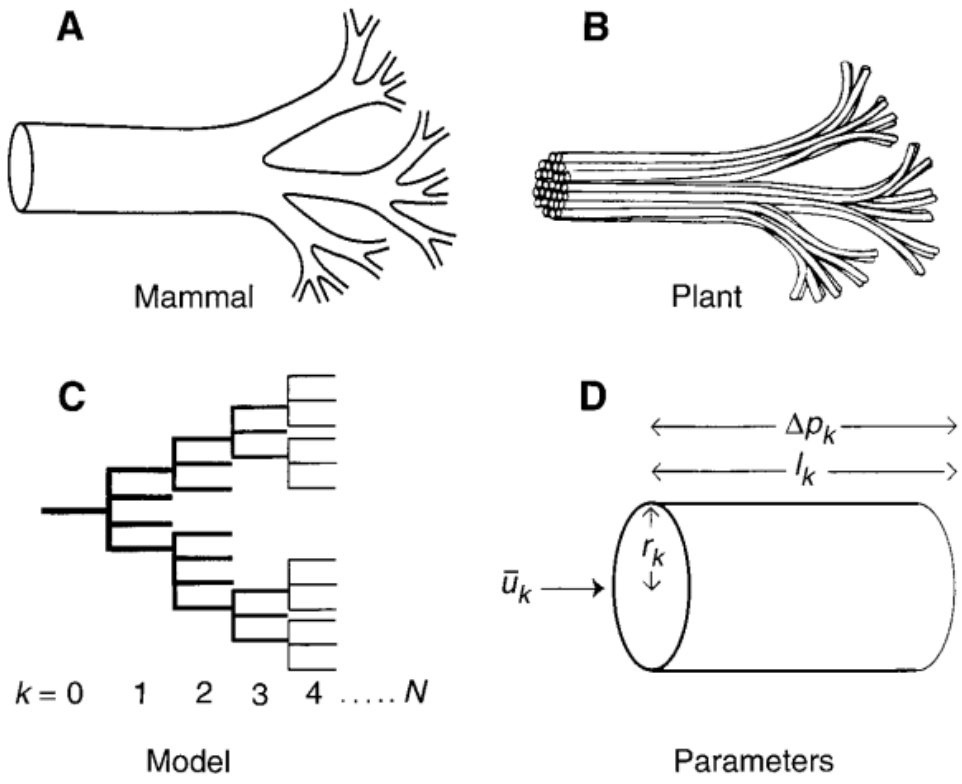
Biology

A general model that describes how essential materials are transported through space-filling fractal networks of branching tubes.

West, Brown, Enquist. 1997 Science

networks

G. West

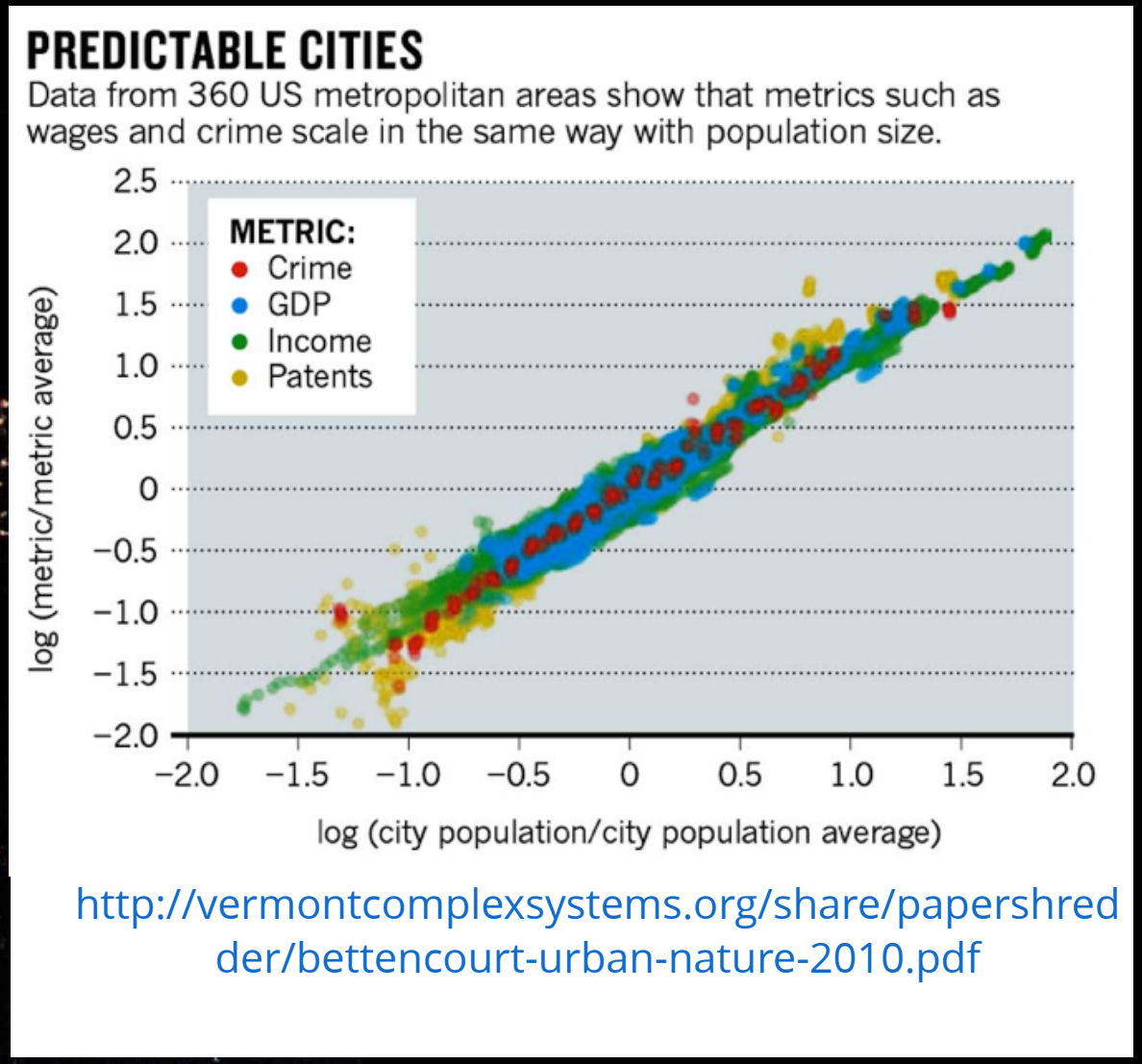


Diagrammatic examples of segments of biological distribution networks

Urban Science

Cities are networks too! And they obey scaling laws on a ridiculus number of parameters!

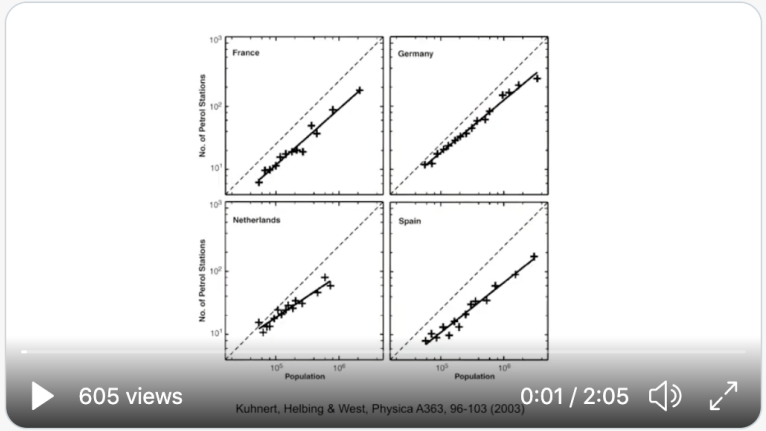
Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Proc. Natl Acad. Sci. USA 104, 7301–7306 (2007)



http://vermontcomplexsystems.org/share/papershredder/bettencourt-urban-nature-2010.pdf

Patrick Sharkey @patrick_sharkey · Sep 22

Yesterday my class on urban inequality discussed Geoffrey West's ideas about "universal" laws of scaling in cities. A few questions and comments came up

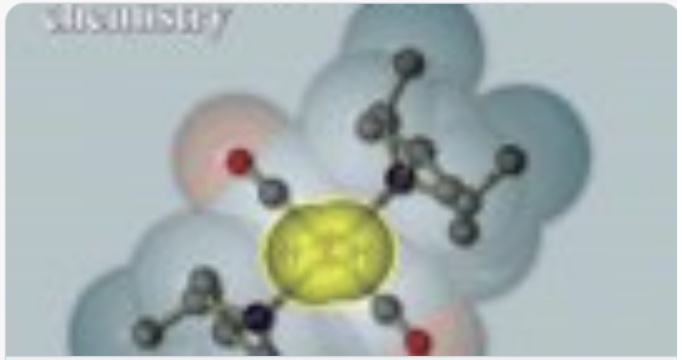


1 3 10

Patrick Sharkey @patrick_sharkey · Sep 22

Replying to @patrick_sharkey

I think the argument is fascinating. But the first q is: is his claim right? His data sources were all over the place, some pretty shaky (e.g. pace of movement!). Has there been work testing his claims?



Growth, innovation, scaling, and the pace of life in cities

Humanity has just crossed a major landmark in its history with the majority of people now living in cities. Cities have long been known to...

pnas.org

1

Patrick Sharkey @patrick_sharkey · Sep 22

For violent crime, eg, the largest cities in the US no longer have the highest levels of violence. Big cities mobilized to confront violence (thru brute force policing, mass incarceration, large-scale surveillance, private security, and *community mobilization*) - violence fell.

1

Patrick Sharkey @patrick_sharkey · Sep 22

Second, I would argue that his "laws" miss the point: growth itself, and the *distribution* of resources, people, disease, crime etc are the crucial phenomena that urbanists care about.

1

Patrick Sharkey @patrick_sharkey · Sep 22

West takes growth for granted - but why do some cities grown and others don't? Which actors play a central role? this is where power, politics, movements, and conflict drive change, where "growth machine" theory enters in, where "homevoters" become relevant.

1

Patrick Sharkey @patrick_sharkey · Sep 22

We read pieces of Urban Fortunes as well, to make the point that the very thing that West takes for granted - growth - is the thing that planners, geographers, political scientists, sociologists consider most crucial to understanding urban inequality and change.

1

Patrick Sharkey @patrick_sharkey · Sep 22

Our key questions are: Where does growth happen? What does it look like? How are resources and people distributed across communities? Which actors and institutions play a central role?

1

Patrick Sharkey @patrick_sharkey · Sep 22

My conclusion: When one tries to ignore all of this and argue for universal laws of all cities, one misses virtually everything important about how cities work and how urban inequality emerges and changes. There's nothing natural, inherent, or inevitable about urban development.

1 5

https://twitter.com/patrick_sharkey/status/1308417626574655488?s=20

key concepts

model residuals - they should be small, they should be random

correlation

correlation vs causation

autocorrelation

spatial correlation

spatial weights

spatial lag

references

https://www.youtube.com/user/GeoDaCenter

LUC ANSELIN'S RECORDED LECTURES ON YOUTUBE

Next wed we will have a ethics of autonomous vehicle class. For that, prepare by reading this.

https://points.datasociety.net/toward-accountability-6096e38878f0

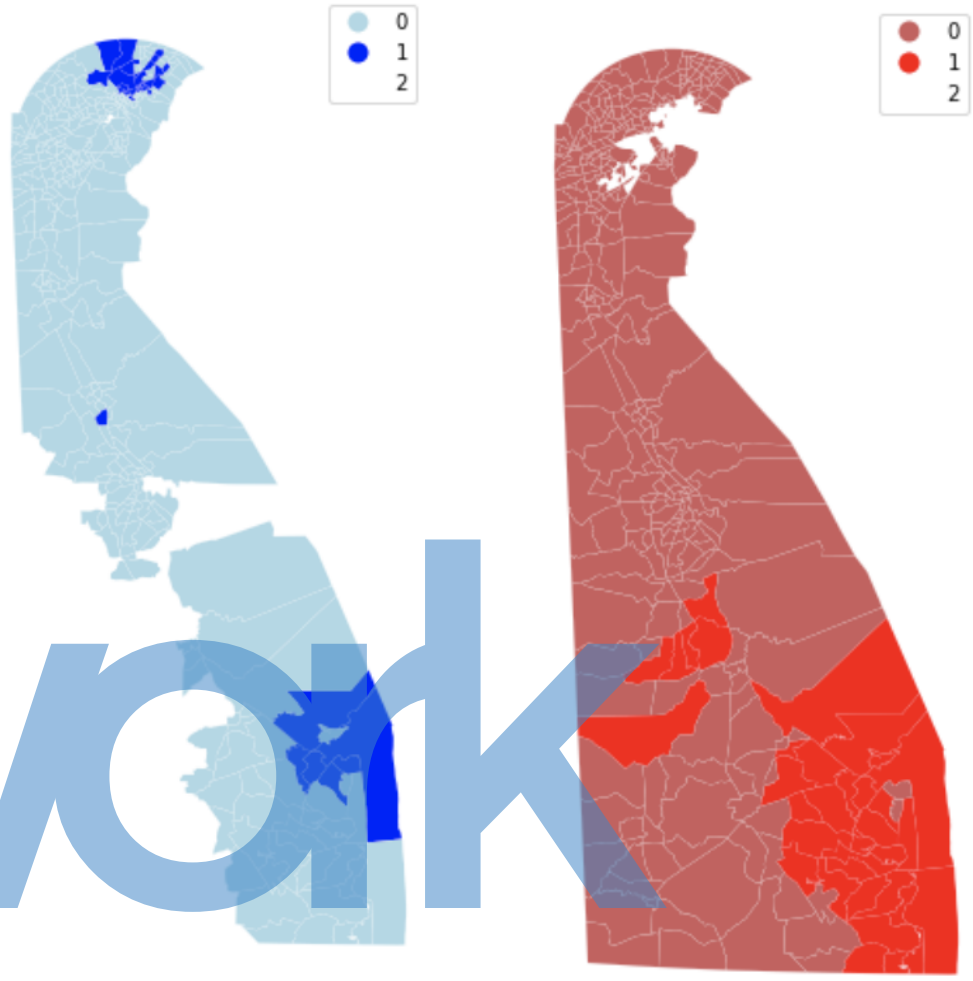
Start from your HW5 (or from the solution posted)

The HW5 solution follows. Some of the tasks to are done differently than I would have if I had only the HW5 tasks to do. Those things are outlined below and they are generally in cells of code that need to be finished. The HW6 tasks proper start below ehre it says "**This is where HW2 starts**"

You will need to rerun it and make dure that

- there are no nan values. Since a lot of the analysis you do relies on population density this will have to be done by **removing areas of 0 population**
- after you remove precints fix the index by calling `df_.reset_index(inplace=True)` on your dataframe so that it does not have missing value. if it did you would be in trouble after creating the weight: you would get a missing index value (I got it for index 274) and also you might have troubles when fitting a line to the Moran's I
- make sure you convert the dataframe to feet (epsg=2263) and work your spatial analysis in feet (there is one point in which I will give you the coordinates of a specific locatoin in lat and lon: you need to either convert those in feet or reconvert temporarily the dataframe to lat-lon to do that part of the analysis)
- make sure you select a single race! the vanilla analysis we did in HW5 could work with summing all the races, but if we want to really look at voters turout, and especially voter turnout by political party and analyze blue and red votes separately, we need to make sure we do not doublecount. I chose the U.S. Senate race. If this were a real project you would need to check if choosing another rate changes your result.

NOTE: I care mostly about the interpretation of the figures in this notebook more than ever! if you cannot get a piece of code to reproduce the figure ask me. I do want you to try (so I want you to ask for it on a step by step basis), but I am happy to show you the solution. I really care about the considerations and interpretation tho!



homework

This is where HW2 starts

feature engineering

- create a "voterturnout" variable: the number of votes per person.
- create a "red votes" variable: the number of republican votes per person.
- create a "blue votes" variable: the number of democrat votes per person.