UNIVERSITY OF CALIFORNIA, SAN DIEGO

THE GUARDIAN

Extra! Extra!

COGS108: Data Science In Practice

Inferential Analysis

GitHub Users Have Desired Job Skills!

Data Science Is The Future

COGS108 Students Blow Professor Away With Interesting Final Projects.

Question

Does prestige increase newspaper readership?

"How would you measure prestige?"

* wikipedia cites each source
  - where cited?
* journalism awards
  - → Pulitzer Prize
* how long in circulation
* where they get $ from

readership - physical copies
  - may not be recicl
- web user content
  - → views
  - → scroll rate
  - → ad revenue (proxy) online
  - → # of subscriptions
    - → miss out on "free" readers

i.e. FB questionable

Question

Data Science Question

prestige → readership

Politzer → "subscription + physical copies"

which newspapers w/ higher "readership metric" have more Politzers?

What time period?
What newspapers?
What specific metrics?

start w/ all newspapers

# of Politzers → mean ± 2SD

time period: paper exist?

reach: national vs local

>5 Politzer → newspapers → # of politzers

"What is the effect of x on y"

Hypothesis

What do you expect the results from this analysis to be?

prestige → readership

→ Prestige ≠ readership
Pulitzers
↳ do readers know how many Pulitzers

→ effect only seen in less read papers; New Yorker vs NYT
"better reading" "lots of readers"

→ will have an effect
opositive, but not super strong
other factors likely playing a larger role

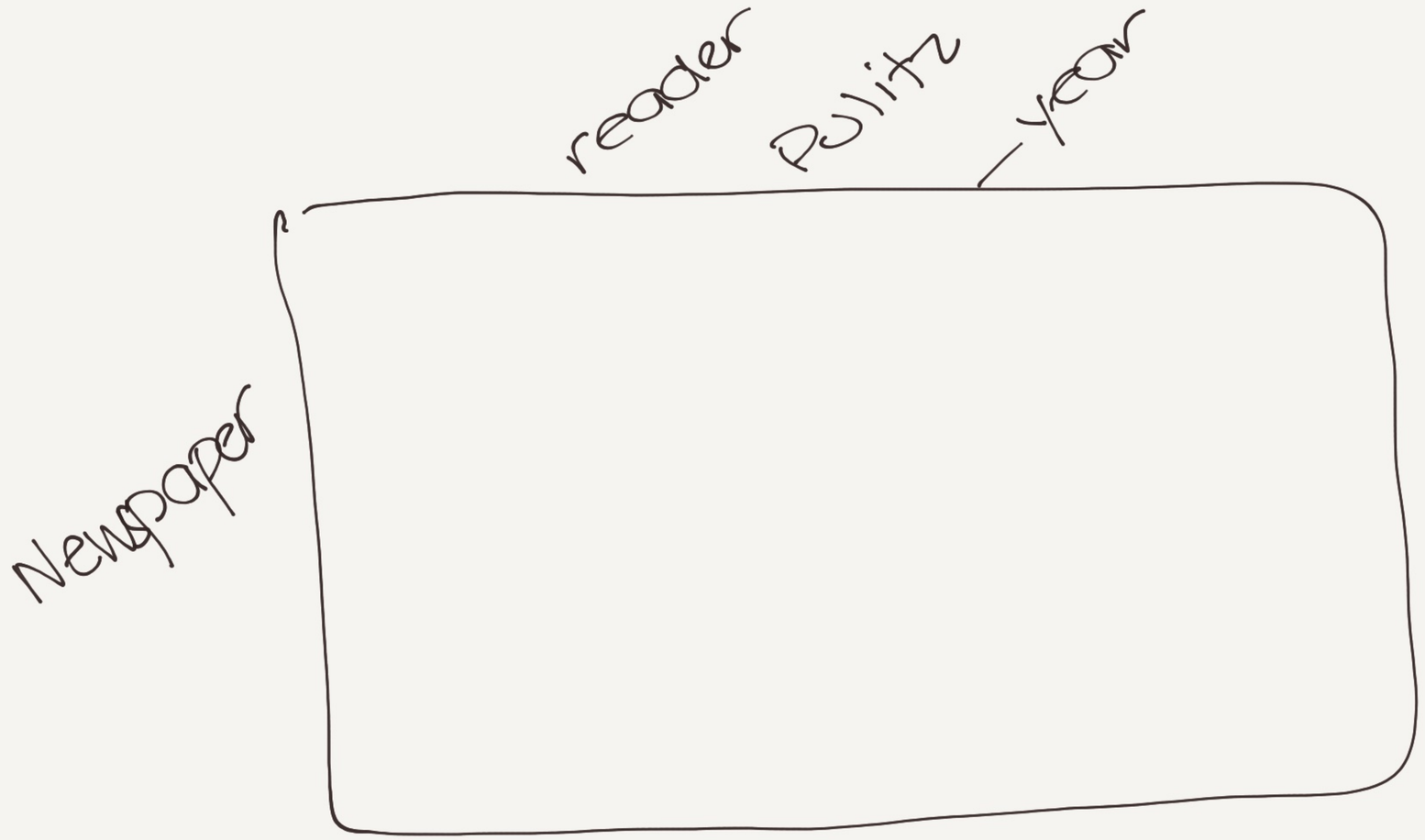
Data

What data/information would you need?

decisions

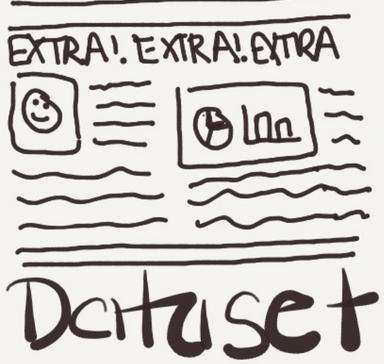
- # of Pulitzers
  - winners + nominees
- measure of read
  - a single year? over time?
- time period
  - internet popular?
- Which newspapers
  - TOP 50?
  - SD from the mean

How would you want this information to be stored? (structured)



What Python tools would you need to use?

- pandas - dataframe
- stats package?
- numpy? - helpful for numbers
- matplotlib - plots
  - pandas
  - seaborn



– 50 rows, 7 columns

Newspaper	2004 Daily Circulation	2013 Daily Circulation	Change in Daily Circulation	Pulitz. 1990-2003	Pulitz 2004-14	Pulitz 1990-2014
USA Today	2,192,098	1,674,306	-24%	1	1	2
WSJ	2,101,017	2,378,827	+13%	30	20	50
NYT	.	.	.	.	.	.
LAT	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Investor's Business Daily	215,735	157,161	-27%	0	1	1

THE GUARDIAN
EXTRA! EXTRA! EXTRA
REVIEW

→ What is the effect of prestige (as measured by Pulitzers won) on newspaper readership?

Q:

H: - no effect
- pos. effect
- readers don't care/know
- other things more important

Data:

50 newspapers

Newspaper	2004 Daily Circulation	2013 Daily Circulation	Change in Daily Circulation	Pulitz. 1990-2003	Pulitz 2004-14	Pulitz 1990-2014
USA Today	2,192,098	1,674,306	-24%	1	1	2
WSJ	2,101,017	2,378,827	+13%	30	20	50
NYT						
LAT						
...						

THE GUARDIAN

EXTRA! EXTRA! EXTRA

How would you explore the data?

EDA

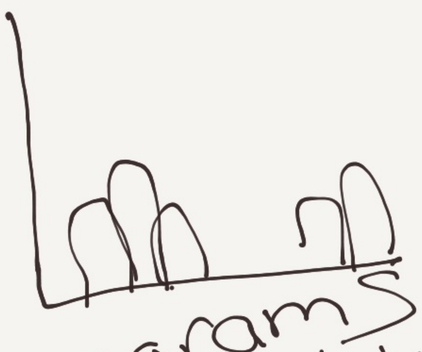
- overall trend
  - daily circulation
  - pulitzer
  - pick year + look
  - scatterplot

- change in Pulitzer + change in circulation

scatterplot

- before relationship
  - check linear regression assumptions
    - normally distr
    - LD transform?

- missingness?
  - * no missing data



histograms
- individual distr

Newspaper	2004 Daily Circulation	2013 Daily Circulation	Change in Daily Circulation	Pulitz. 1990-2003	Pulitz 2004-14	Pulitz 1990-2014
USA Today	2,192,098	1,674,306	-24%	1	1	2
WSJ	2,101,017	2,378,827	+13%	30	20	50

EDA

what exploratory visualizations would you generate?

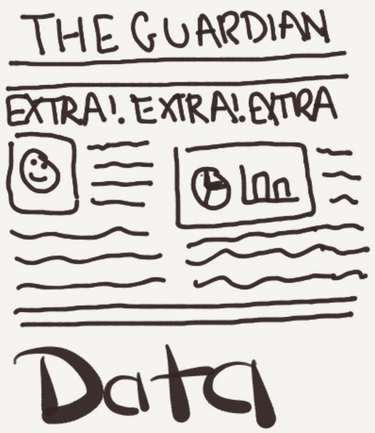
- check extremes
  - really high/low?

Newspaper	2004 Daily Circulation	2013 Daily Circulation	Change in Daily Circulation	Pulitz. 1990-2003	Pulitz 2004-14	Pulitz 1990-2014
USA Today	2,192,098	1,674,306	-24%	1	1	2
WSJ	2,101,017	2,378,827	+13%	30	20	50

EDA

what exploratory visualizations would you generate?

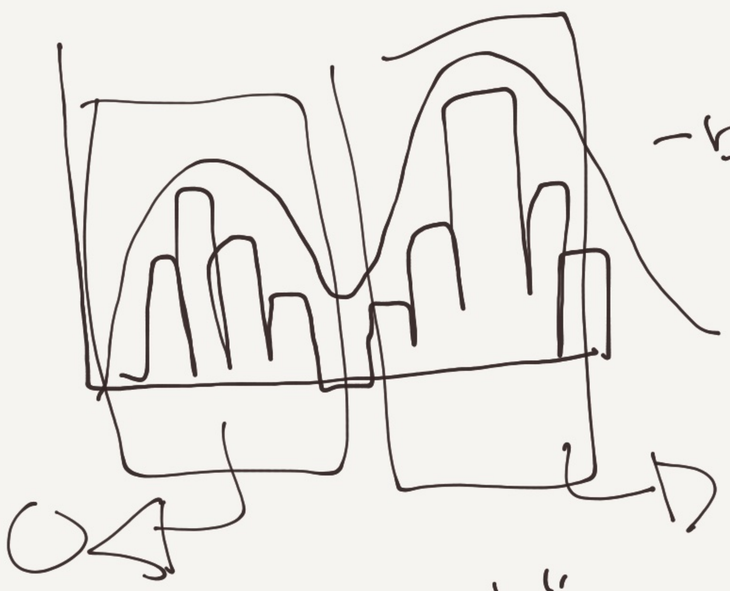
Newspaper	2004 Daily Circulation	2013 Daily Circulation	Change in Daily Circulation	Pulitz. 1990-2003	Pulitz 2004-14	Pulitz 1990-2014
USA Today	2,192,098	1,674,306	-24%	1	1	2
WSJ	2,101,017	2,378,827	+13%	30	20	50



Data

What if our data didn't look as we anticipated?

- transformations
  - log
  - $x^2; 3:$
  -
- outliers
- non-normal distributions



- bimodal
  - ↳ ⇒ stratification
  - ↳ "recode"

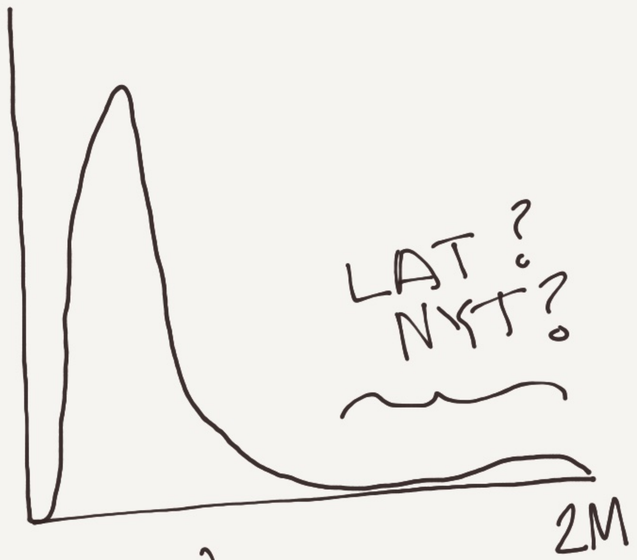
↳ check normal
test for normality

⇓

interpretation

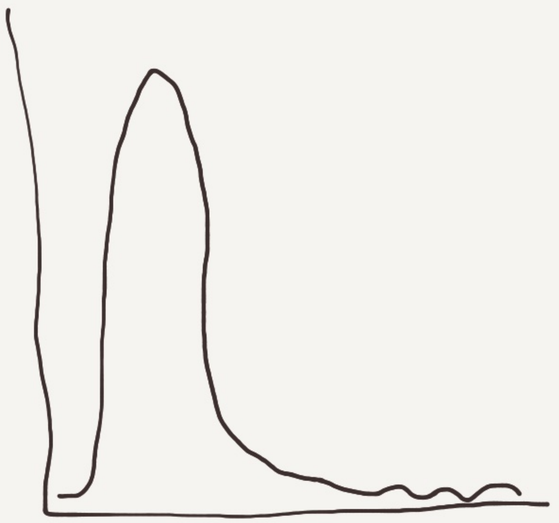
EDA: Circulation (Readership)

Daily Circulation 2004

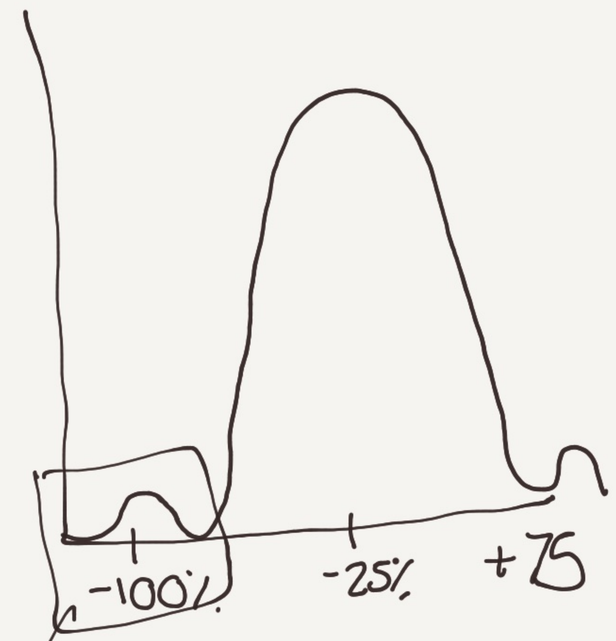


lots of publications w/ lower circulation

Daily Circulation 2013



Change in Daily Circulation 2004-2013

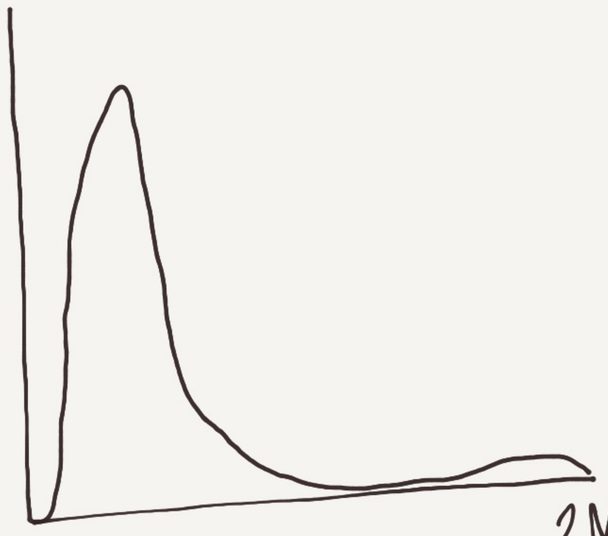


% change

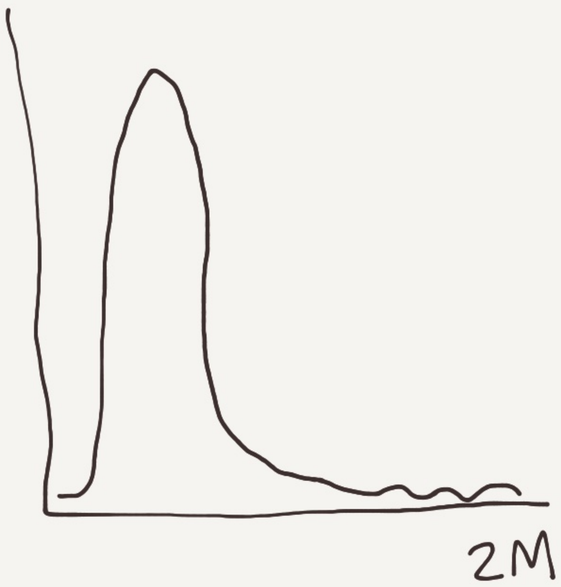
Outliers
- Out of business
- Out of daily circulation

EDA: Circulation (Readership)

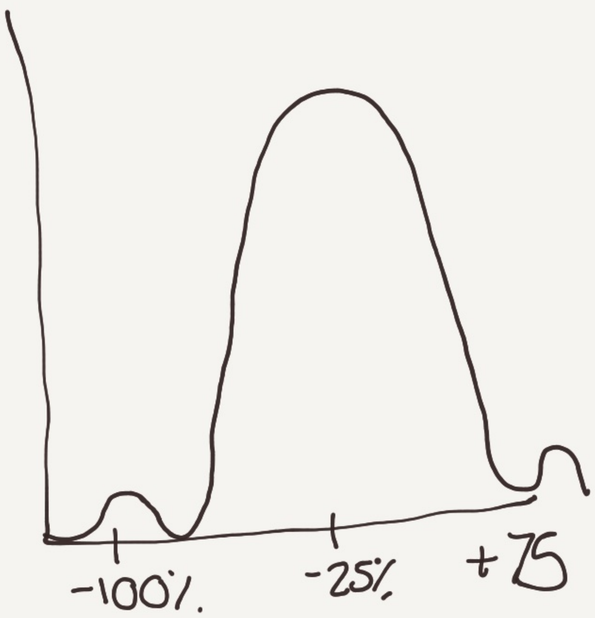
Daily Circulation 2004



Daily Circulation 2013



Change in Daily Circulation 2004-2013



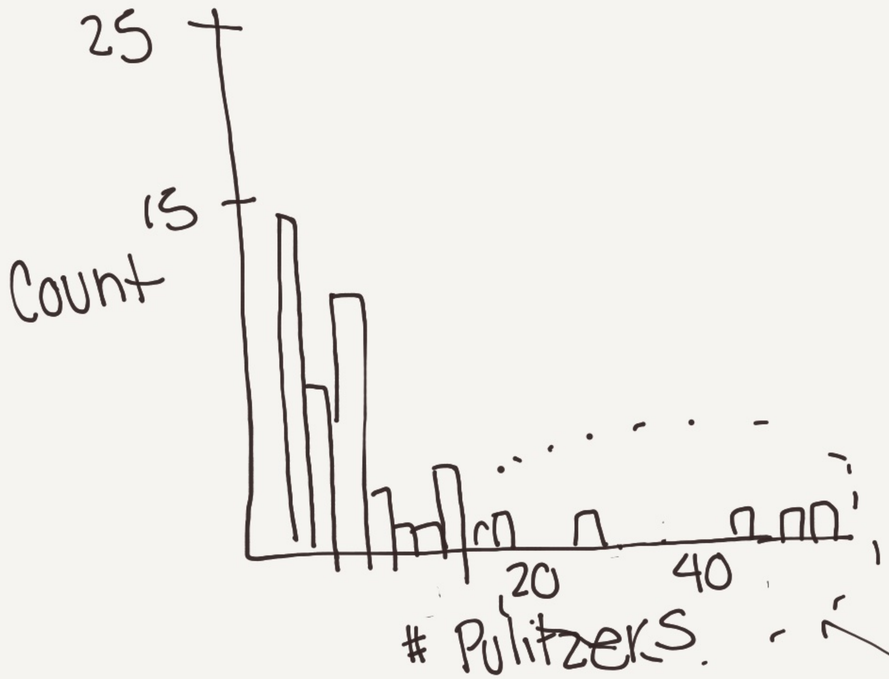
Clicker Question: Which variable would you use for this analysis?

☆ A Daily Circulation, 2004 (18%)
B Daily Circulation, 2013 (17%)
C Change in Daily Circulation (64%)

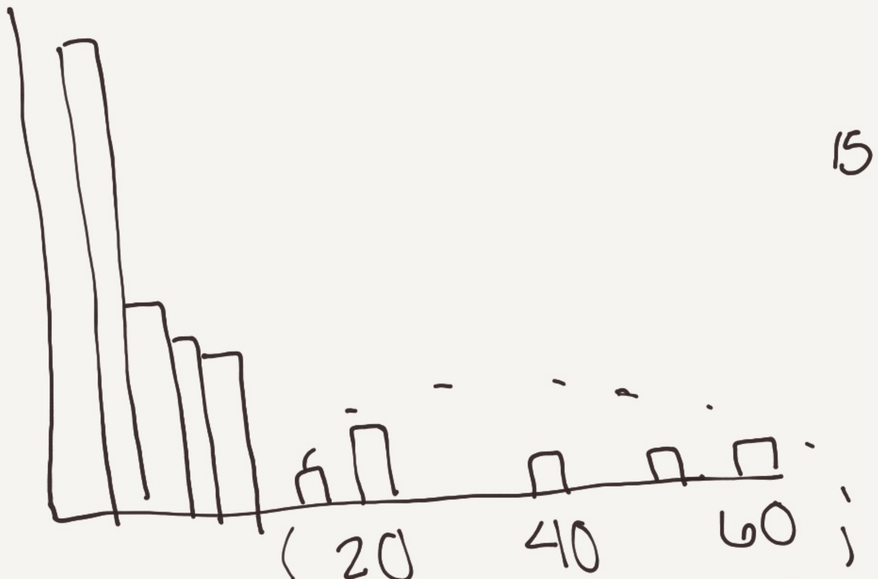
EDA : Pulitzers

Lo # Winners + Finalists

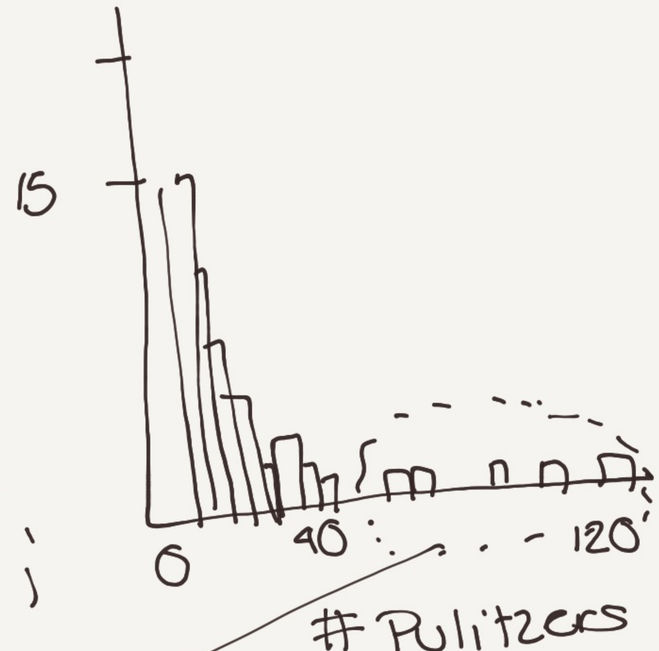
1990-2003



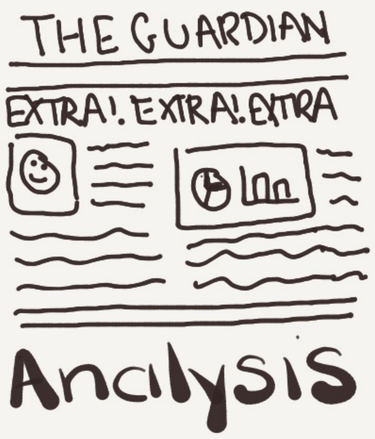
2004-2014



1990-2014



a few publications get way more Pulitzers



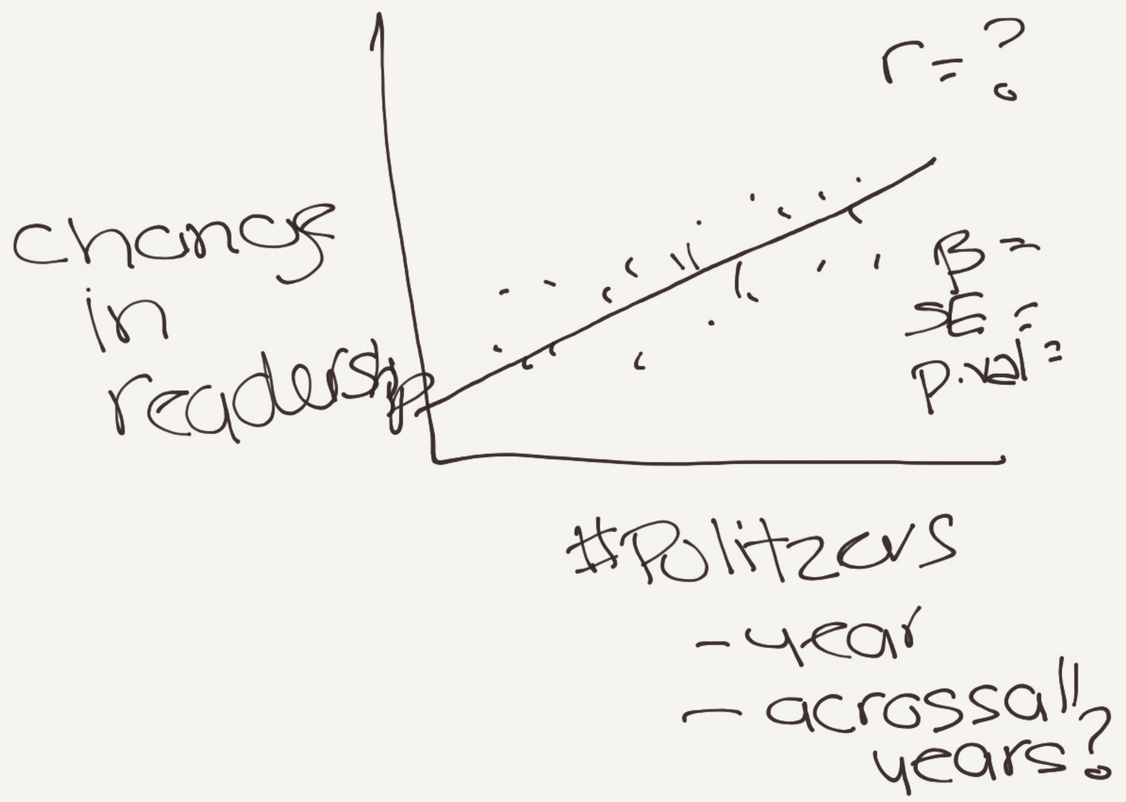
How would you answer the question?

what the effect?

prestige (Politzers) ⟶ readership

H_0: There is no effect (correlation)

H_a: There is an effect (correlation)



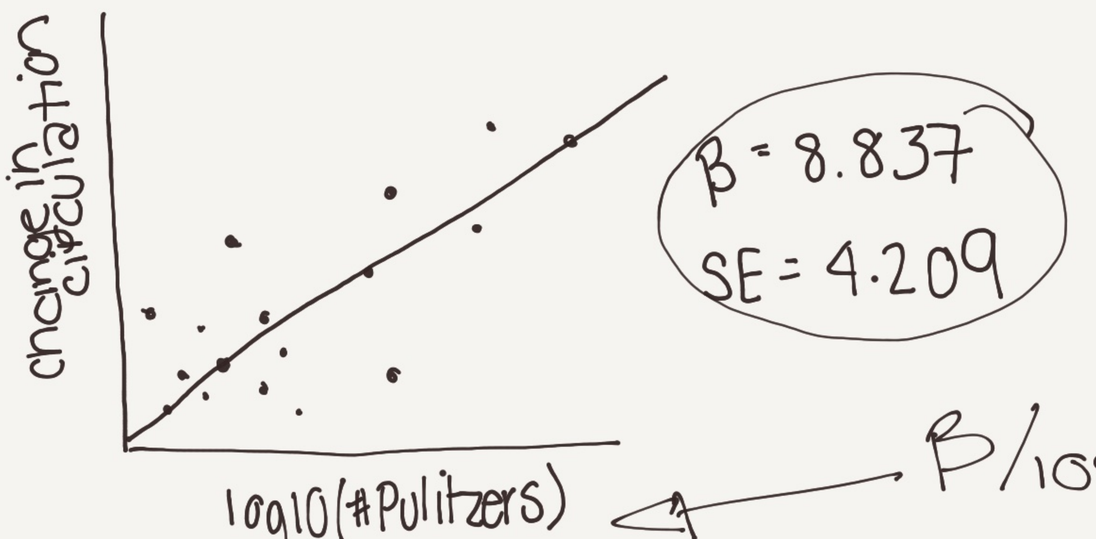
THE GUARDIAN

EXTRA! EXTRA! EXTRA

INTERPRETATION

CLICKER

GIVEN THE FOLLOWING RESULTS, WHAT WOULD YOU CONCLUDE?



-A; SE looks large

B/100 ≈ 0.08 ↑

(70%) A. Prestige increases circulation
(12%) B. Prestige decreases circulation
(7%) C. Prestige does not affect circulation
(11%) D. Something else

↳ correlation

What limitations are there in this analysis?

- small dataset
- ommitted variables
  - quality of newspapers
  - confounders?
- other prestige?
  - citations