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?

prestige?

- how long in circulation?
  - → older = more prestige
- (- blocked in other countries)
- (- hate they receive?)

→ Pulitzer prize winners writing for the newspaper

→ circulation?

(→ do we care about readership ability) want to know
- free time?

→ price of subscription ← proxy
- ↑ prestige ↑ $/md

→ revenue
- ↑ revenue ↑ influence / prestige

→ cited by others / in general
- ↑ citations ↑ prestigious

readership

→ # of readers

(→ type of readers)

→ online
- how long spent?
- scroll rate

→ circulation?
- (- just recycle)
- - multiple / paper

} est.

→ online?
- subscription
- free loaders

Question

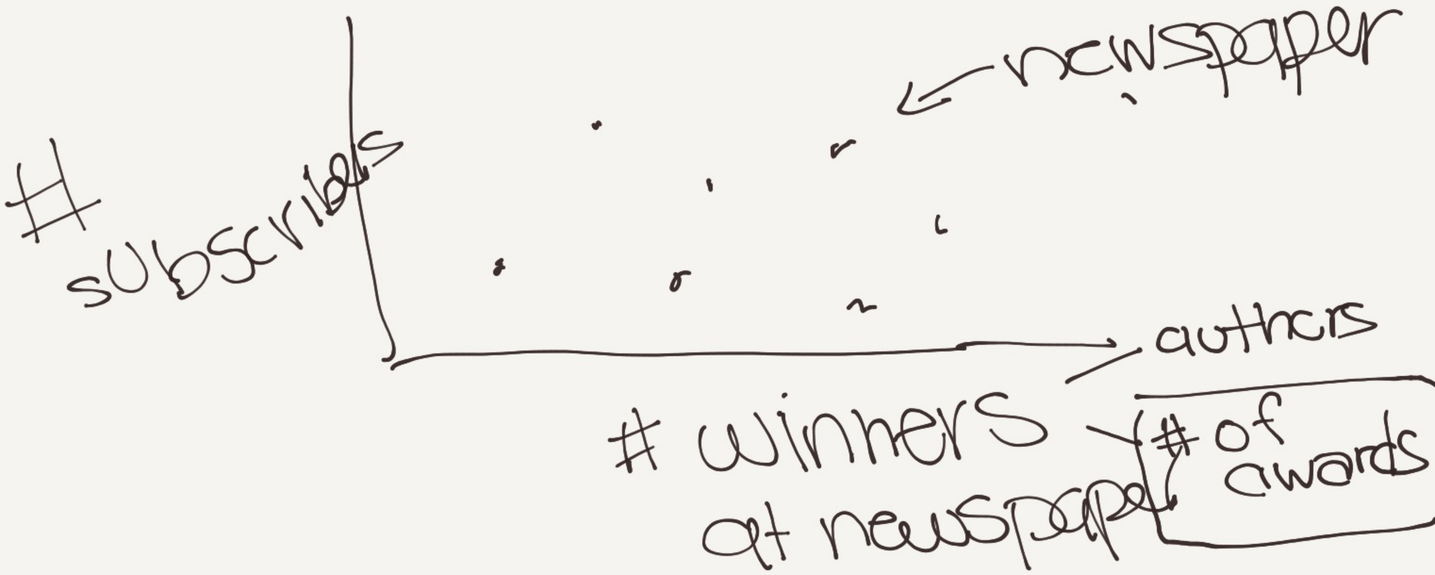
Data Science Question

pulitzer → circulation?

Which papers?
What time period?
What measured?

How many pulitzer prize winners at a newspaper affect # of subscribers?

What is the effect of the number of pulitzers a newspaper receives on it's # of readers



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

Hypothesis

- ↑winners ↑subscribers → Pulitzer = better/more interesting → people would want to read

pulitzers → circulation

quality of writing ← confander

Popular institutions → circulation may NOT explain pulitzers - other things could be at play

- No effect → other things at play - how extreme clickbait → readers don't know about Pulitzers

Data

What data/information would you need?

decisions

→ circulation
(paper/internet)
→ now?

→ # of pulitzers
→ nominees + winners

→ Newspapers
→ TOP 15 in country

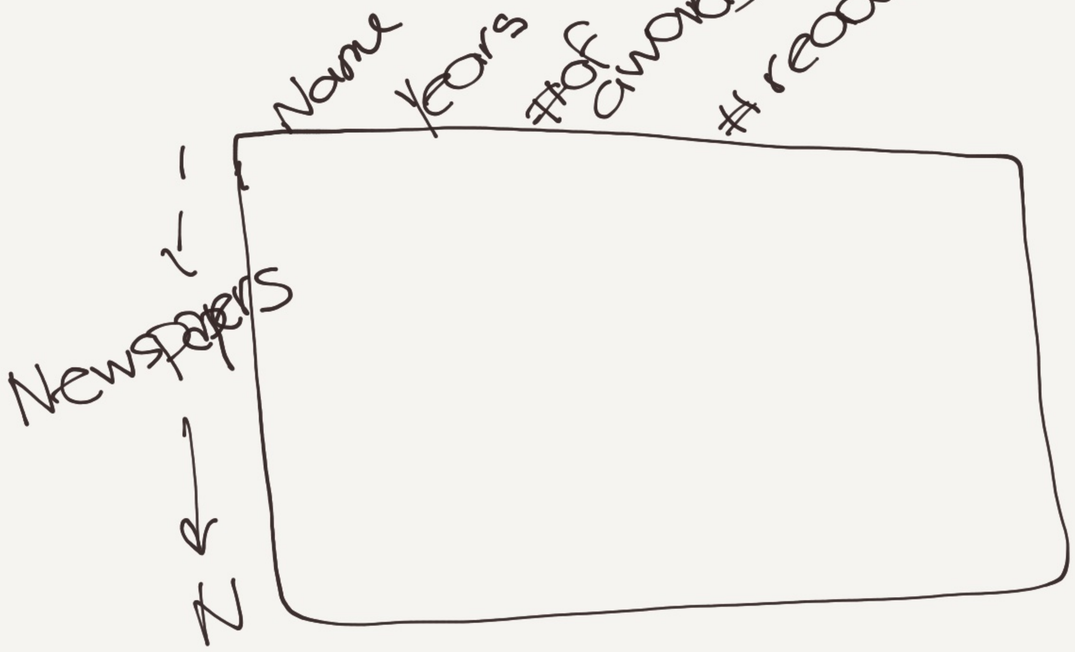
→ years including
* event?

- any newspaper?
w/ pulitzer
- at least 5?
- $\bar{X}$ pulitzers
$\bar{X} - 2*SD$?

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

(- recent years
- prizes)

CSV
- top 10
- across years



← observations in row
variables in columns

What Python tools would you need to use?

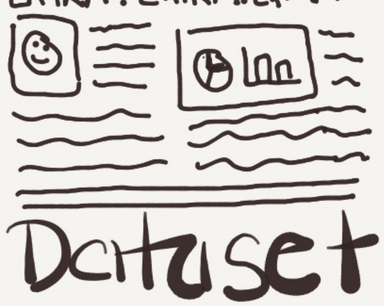
- pandas - Dataframe

- BeautifulSoup (web scraping)

- seaborn (graphs)
  - → matplotlib
  - → pandas

→ Do the analysis?
- correlation - pandas
- numpy (np)

→ stats / statistics package



– 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
- readers don't care/know
- other things more important

- pos. effect

Data:

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						
⋮						

EDA

How would you explore the data?

? magazines

✓ check for outliers?

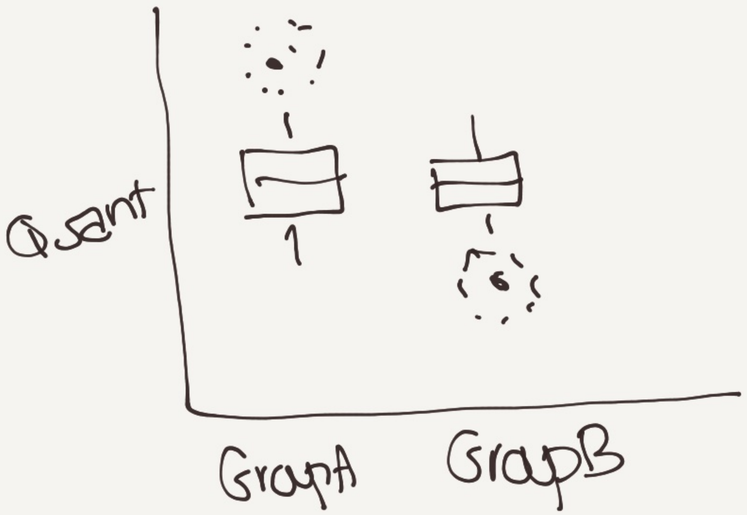
① SD from the mean → * get rid of them?

? ↑ circulation ↑ politzer prizes

✓ - check for missing/null data (No missing data)

Boxplot?

②



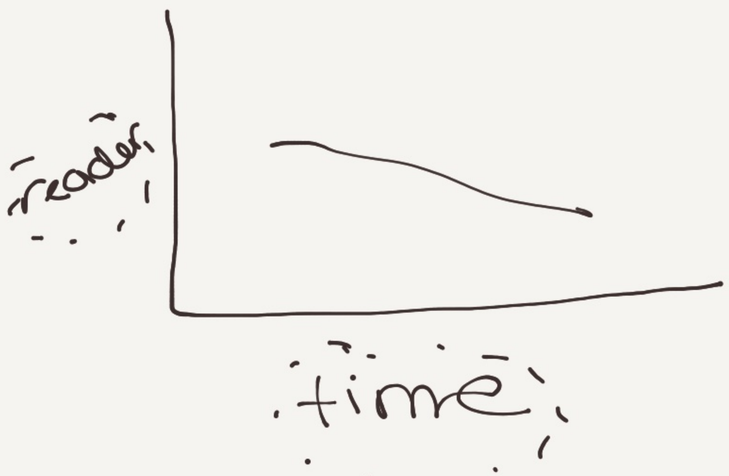
③ outlier

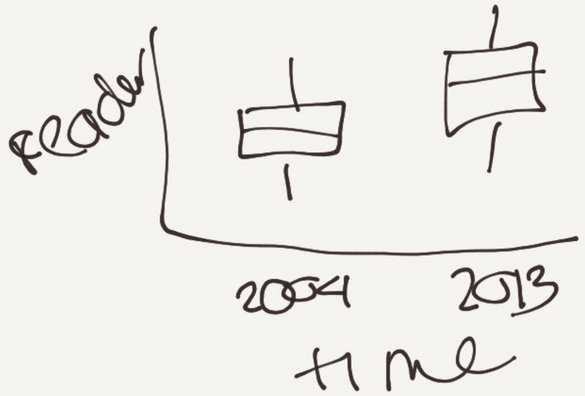
- correlation
- regression
- relationship b/w 2 var.

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

(charge)
increase in
readership
over time





what exploratory visualizations would you generate?

- single variable
  - histograms for each column
    - new papers?
  - density plot

- boxplots } two variables
- scatterplots

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

THE GUARDIAN

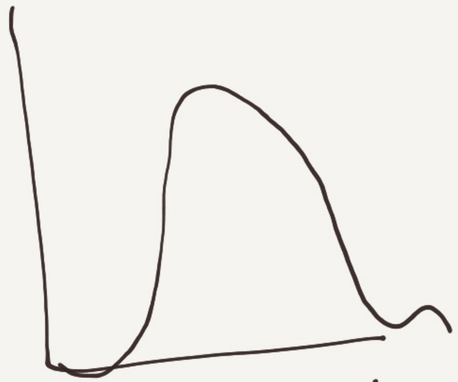
EXTRA! EXTRA! EXTRA

Data

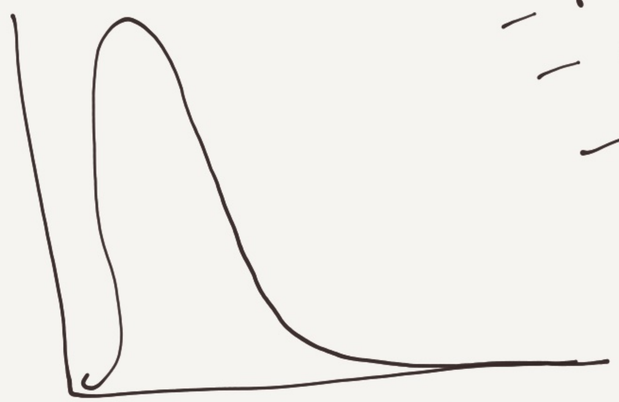
What if our data didn't look as we anticipated?

assumptions

X - multicollinearity
- homoskedasticity
- ~Normal
-



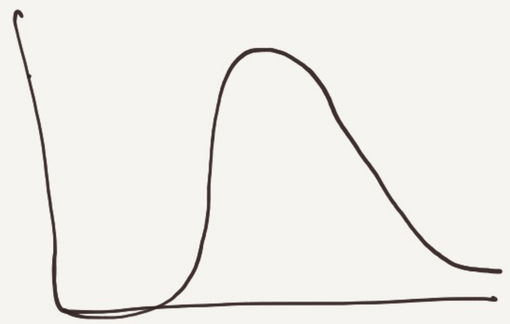
Change in readership
~Normal



Pulitzers
~skewed

stratify

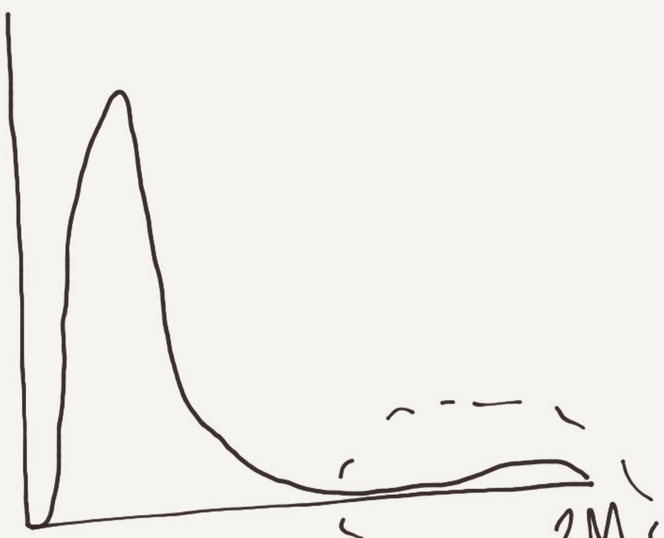
log-transform



EDA: Circulation (Readership)

trend?
interpretation?
looks weird?

Daily Circulation 2004

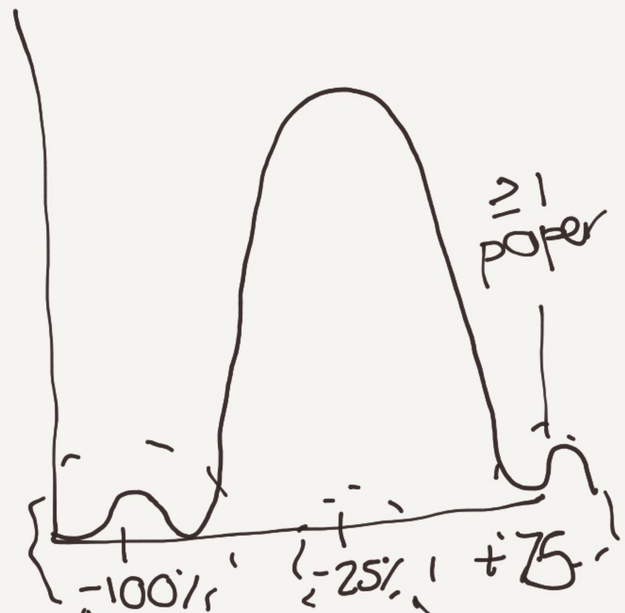


most data for newspapers close to zero

outliers?

Daily Circulation 2013

Change in Daily Circulation 2004-2013



% change

- internet?
- stopped publishing
- No longer daily

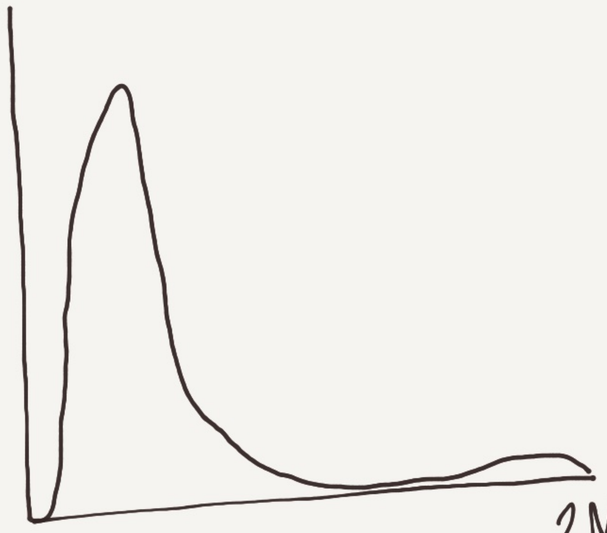
-25? overall readership drop

EDA

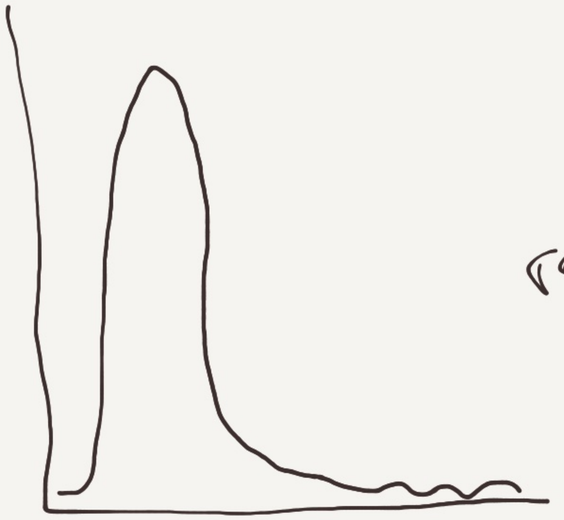
EDA: Circulation (Readership)

prestige → readership

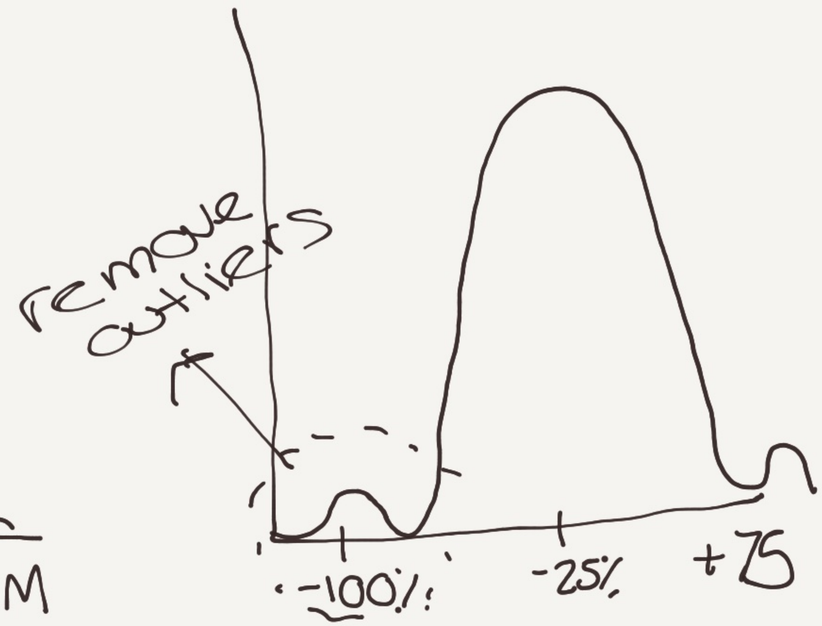
Daily Circulation 2004



Daily Circulation 2013



Change in Daily Circulation 2004-2013



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

☆ A Daily Circulation, 2004 (27)
B Daily Circulation, 2013 (13)
C Change in Daily Circulation (66)

∴ Change

↓

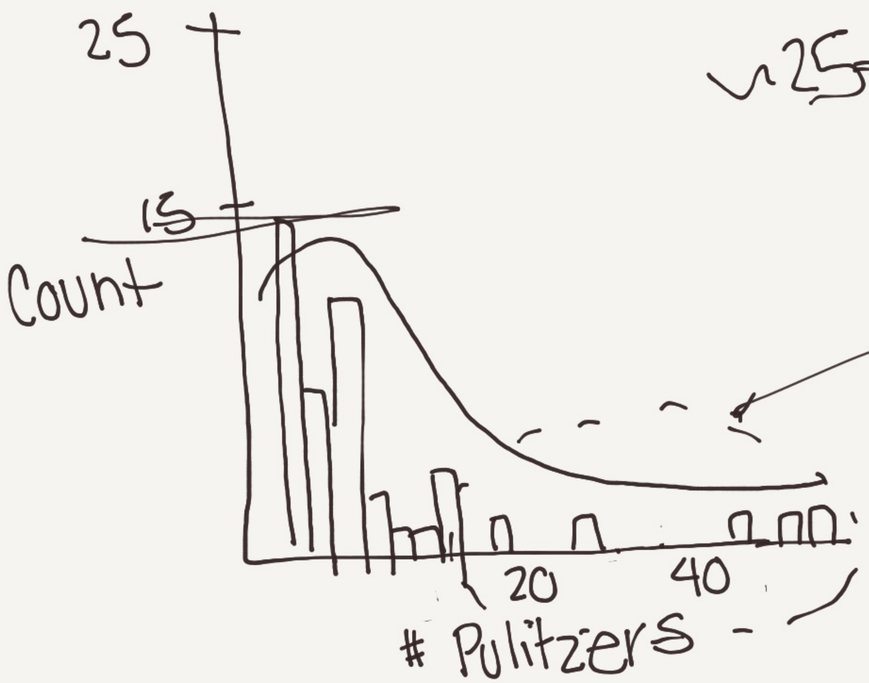
Normal distribution

THE GUARDIAN
EXTRA! EXTRA! EXTRA

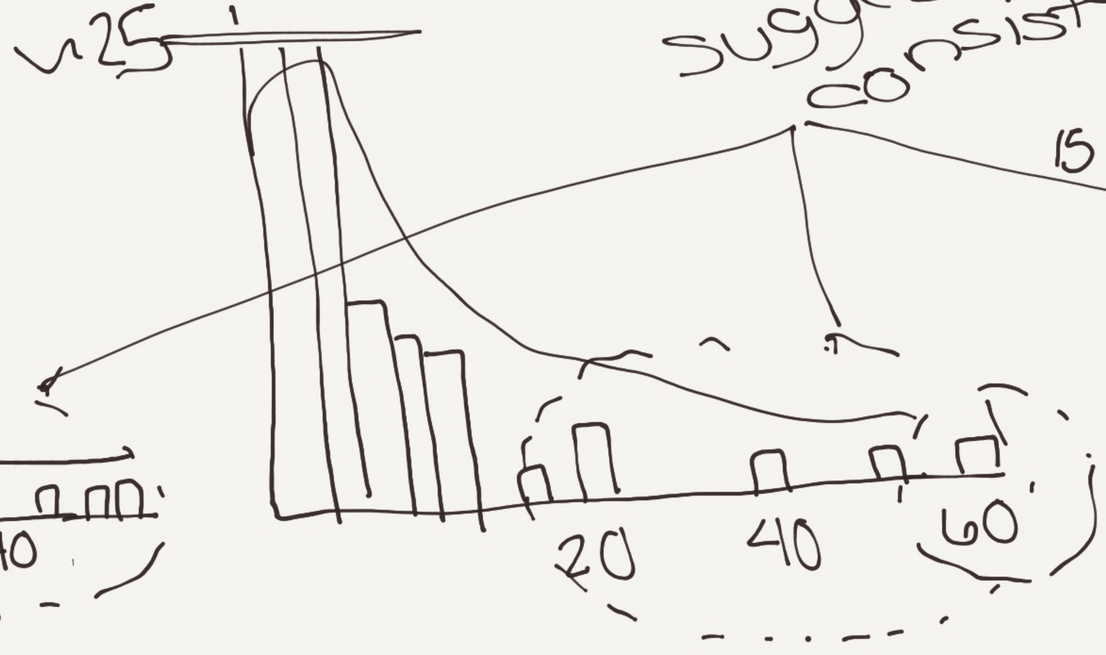
EDA : Pulitzers

↳ # Winners + Finalists

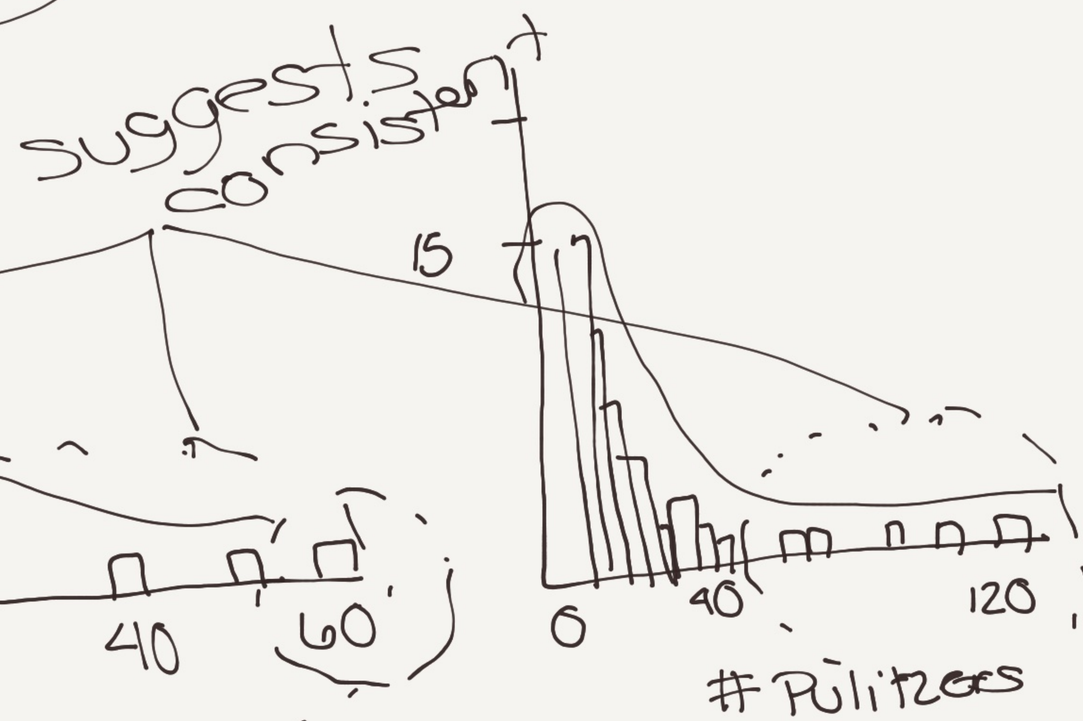
1990-2003



2004-2014



1990-2014



← values closer to lower limit (skew)

↳ more outlets?
↑ news

new award categories? → Pulitzers

THE GUARDIAN

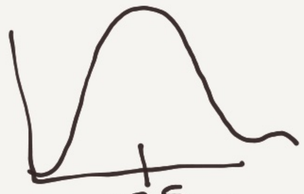
EXTRA! EXTRA! EXTRA

Analysis

How would you answer the question?

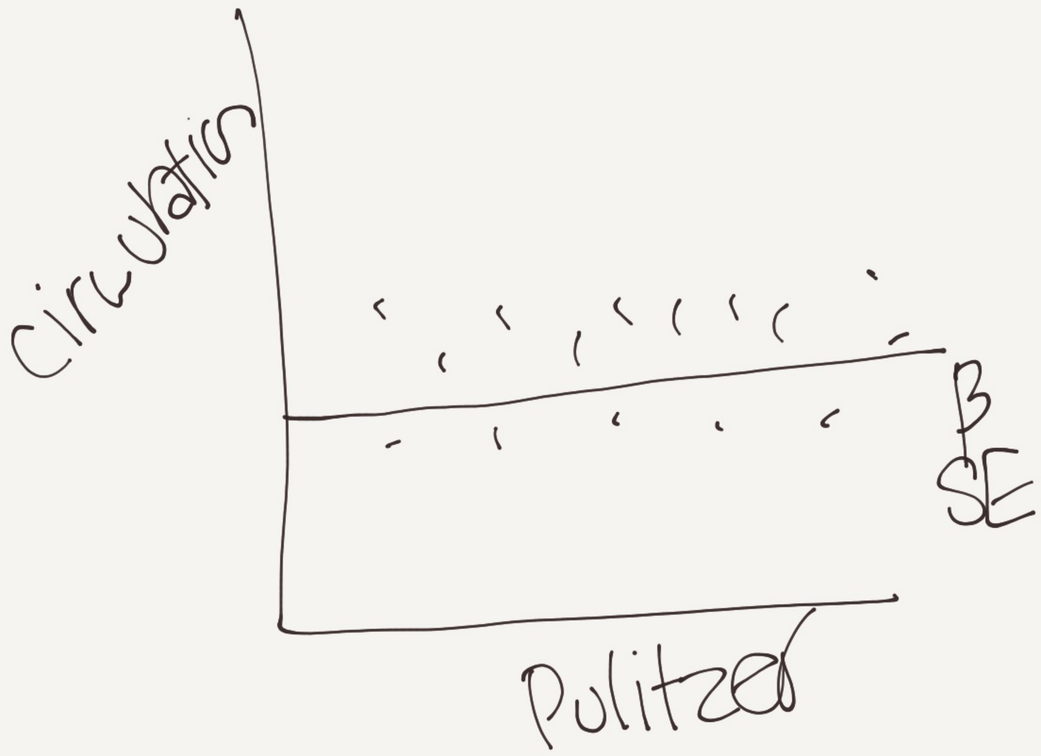
prestige → readership

$\log_{10}$ Pulitzers



-2S

change in circulation



outcome: readership
predictor: pulitzer prestige

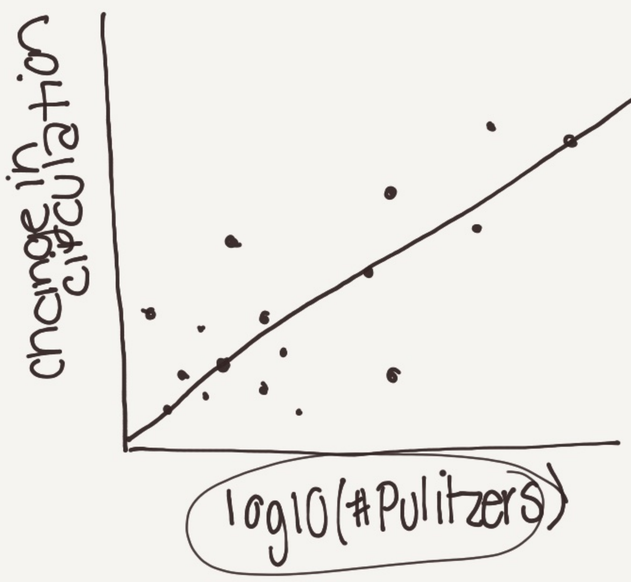
THE GUARDIAN

EXTRA! EXTRA! EXTRA

INTERPRETATION

CLICKER

GIVEN THE FOLLOWING RESULTS, WHAT WOULD YOU CONCLUDE?



$\beta = 8.837$

$SE = 4.209$

$r =$

p.value >=

CI =

$\beta/100 \sim 0.08$! for every Pulitzer change in circul.

An increase in prestige tends to increase/decrease circulation/inc.

82 A. Prestige increases circulation/inc.

5 B. Prestige decreases circulation

9 C. Prestige does not affect circulation

7 D. Something else

↳ correlation

What limitations are there in this analysis?

pre → change in
#Pulitzers circulation

- account for other variables (confounders)
  - ↓ in newspaper (internet measurements affect analysis
- prestige?
  - other measures
  - citations
- measures of readership
- line graph
  → year ... missing