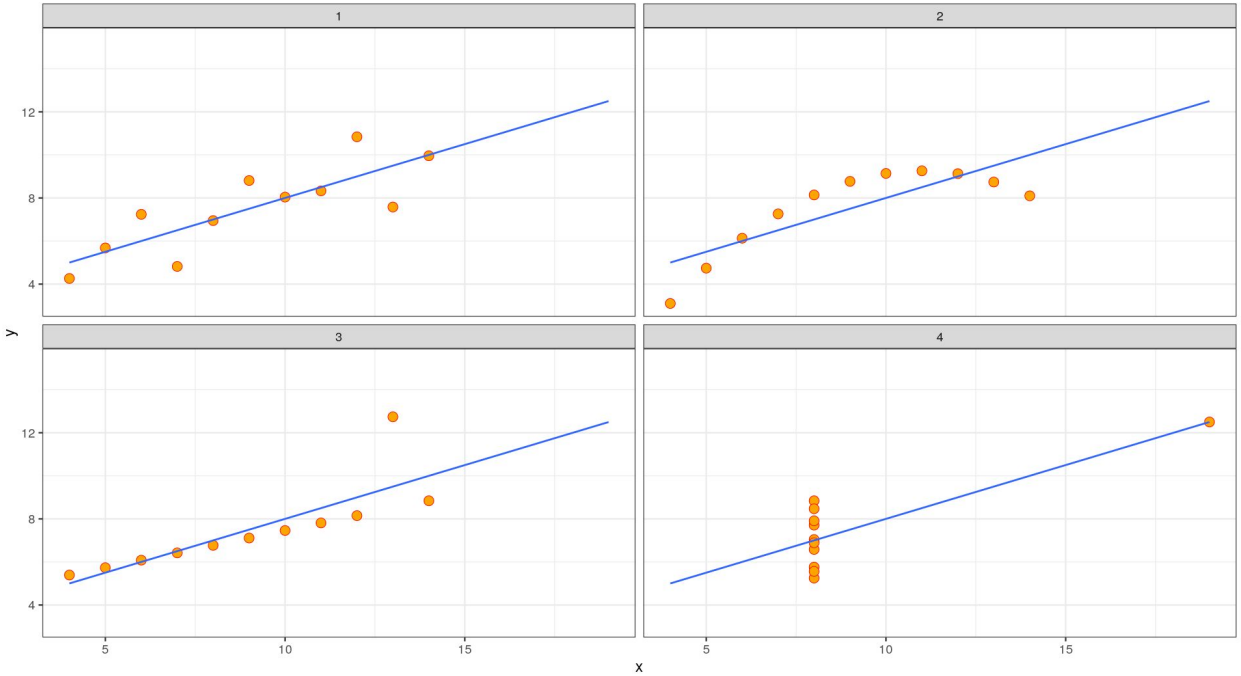
Justin Matejka, George Fitzmaurice (2017): *Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing*

Daróczi Henriett, Környezettudományi Doktori Iskola

2019. október 9.

Anscombe’s Quartet



mean(x)	9.00
mean(y)	7.50
sd(x)	3.32
sd(y)	2.03
cor(x, y)	0.816
lm(y~x)	3 + 0.5x

Source: https://gist.github.com/amoeba/7576126

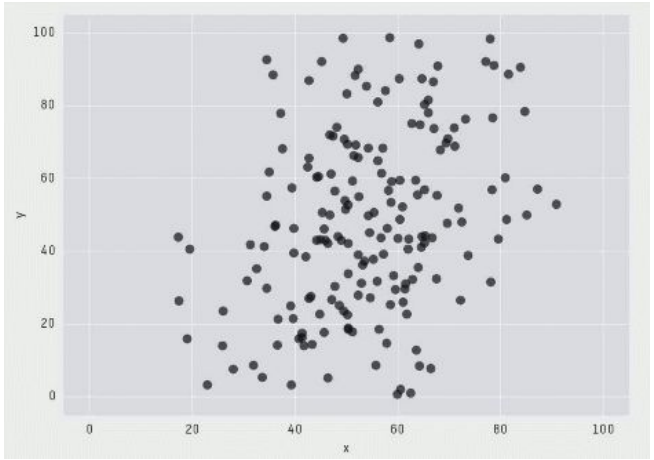
Related Work

- Anscombe (1973): dataset creation method not known
- Chatterjee and Firat (2007): generic algorithm to max graphical dissimilarity
- Govindaraju and Haslett (2009): cloned datasets
- Bach, Spritzer, Lutton, Fekete (2012): random network graphs matching params
- Stefanski (2007): scatter plots to encode graphics

New Results

Generate new datasets with identical statistical properties, but different graphical representation through simulation.

Iterative Algorithm

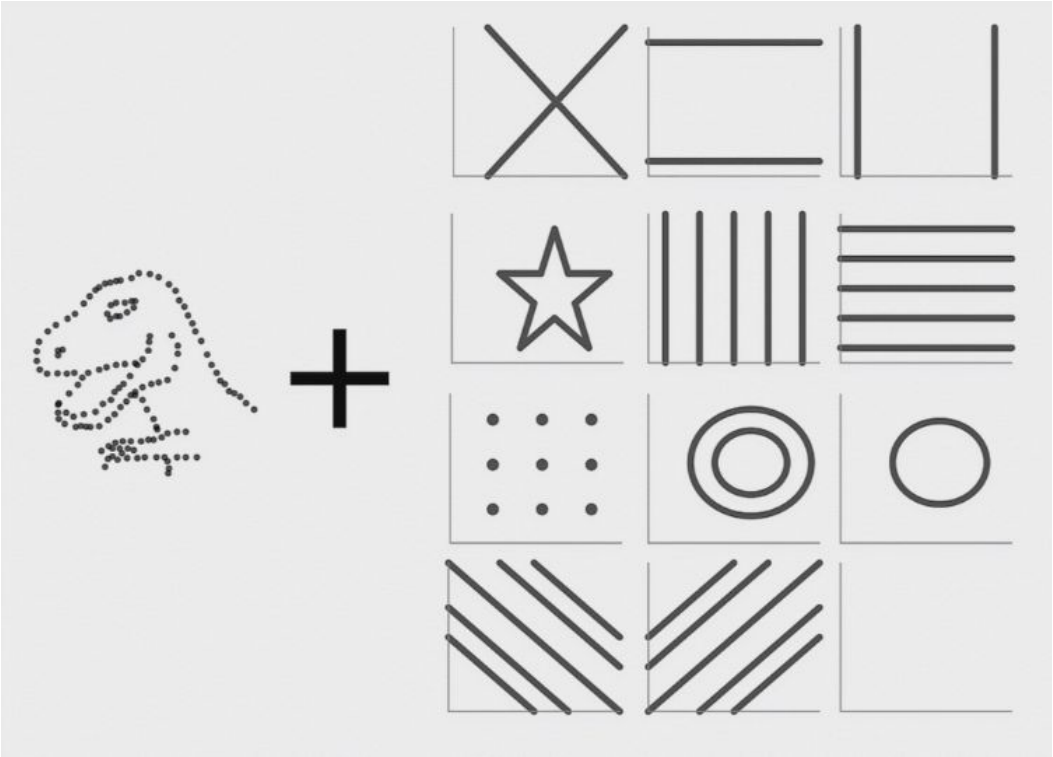


X Mean:	54.0236753
Y Mean:	48.0970794
X SD :	14.5298540
Y SD :	24.7943127
Corr. :	+0.3280926

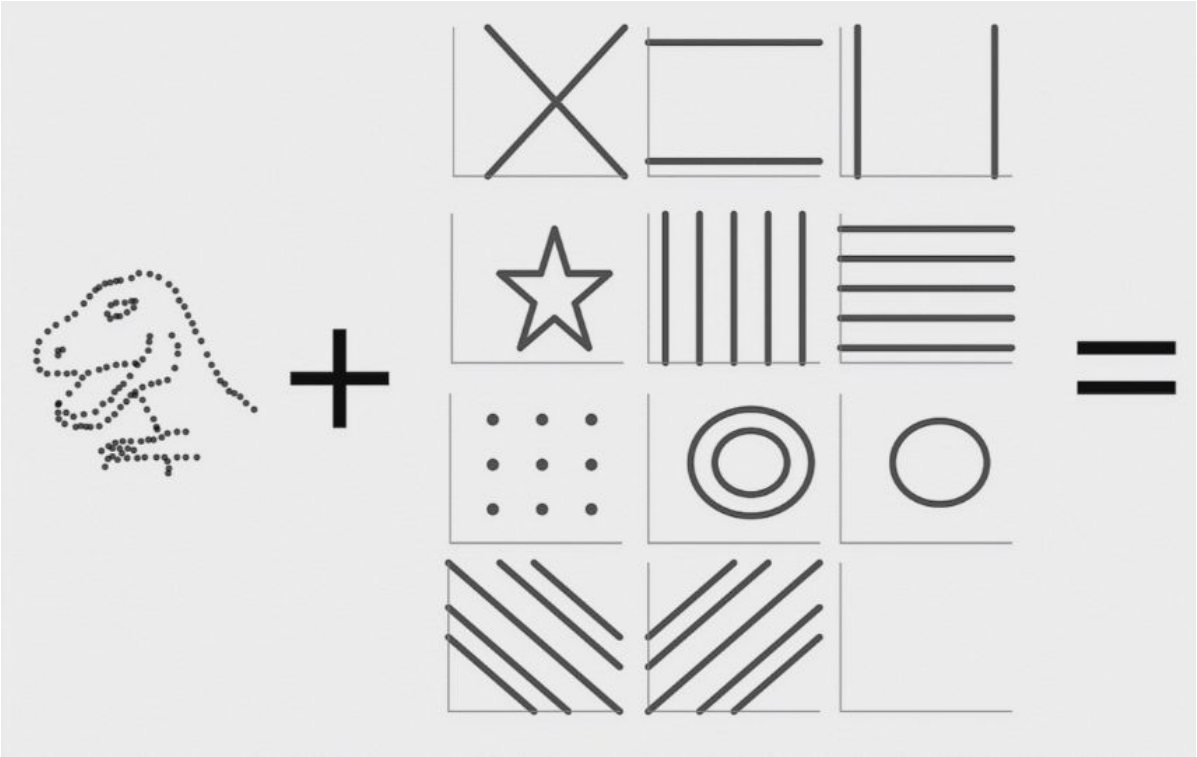
```
1:  current_ds ← initial_ds
2:  for x iterations, do:
3:      test_ds ← PERTURB(current_ds, temp)
4:      if ISERROROK(test_ds, initial_ds):
5:          current_ds ← test_ds
6:
7:  function PERTURB(ds, temp):
8:      loop:
9:          test ← MOVERANDOMPOINTS(ds)
10:         if FIT(test) > FIT(ds) or temp > RANDOM():
11:             return test
```

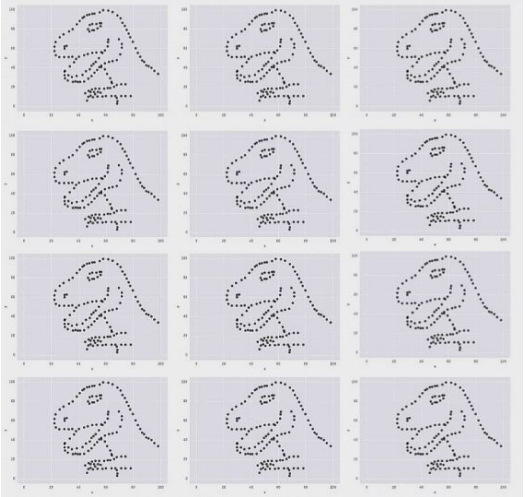
Source: https://www.autodeskresearch.com/publications/samestats

Example Datasets (Alberto Cairo: Datasaurus)

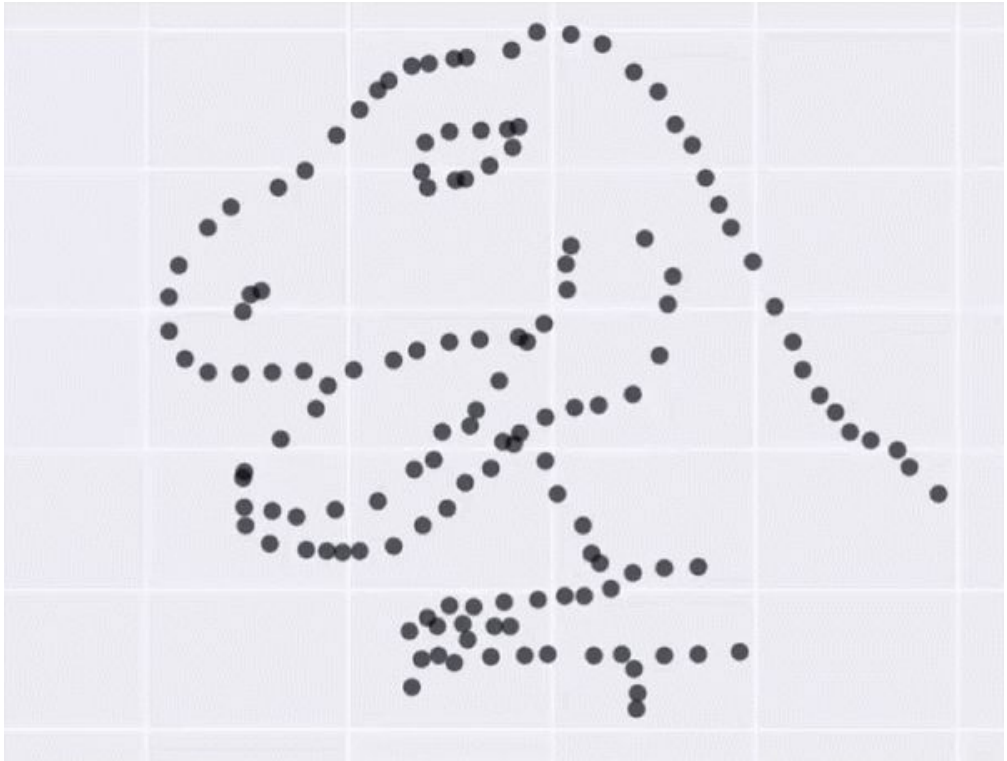


Example Datasets (Alberto Cairo: Datasaurus)





Example Datasets (Alberto Cairo: Datasaurus)



X Mean:	54.26
Y Mean:	47.83
X SD :	16.76
Y SD :	26.93
Corr. :	-0.06

Example Datasets (Simpson's Paradox)

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

Bickel (1975): Sex Bias in Graduate Admissions: Data From Berkeley

Example Datasets (Simpson's Paradox)

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	35%

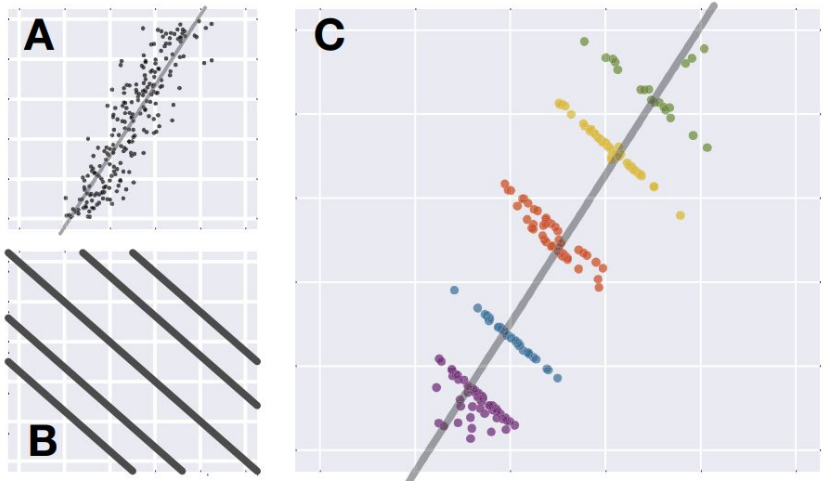
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

Bickel (1975): Sex Bias in Graduate Admissions: Data From Berkeley

Example Datasets (Simpson's Paradox)

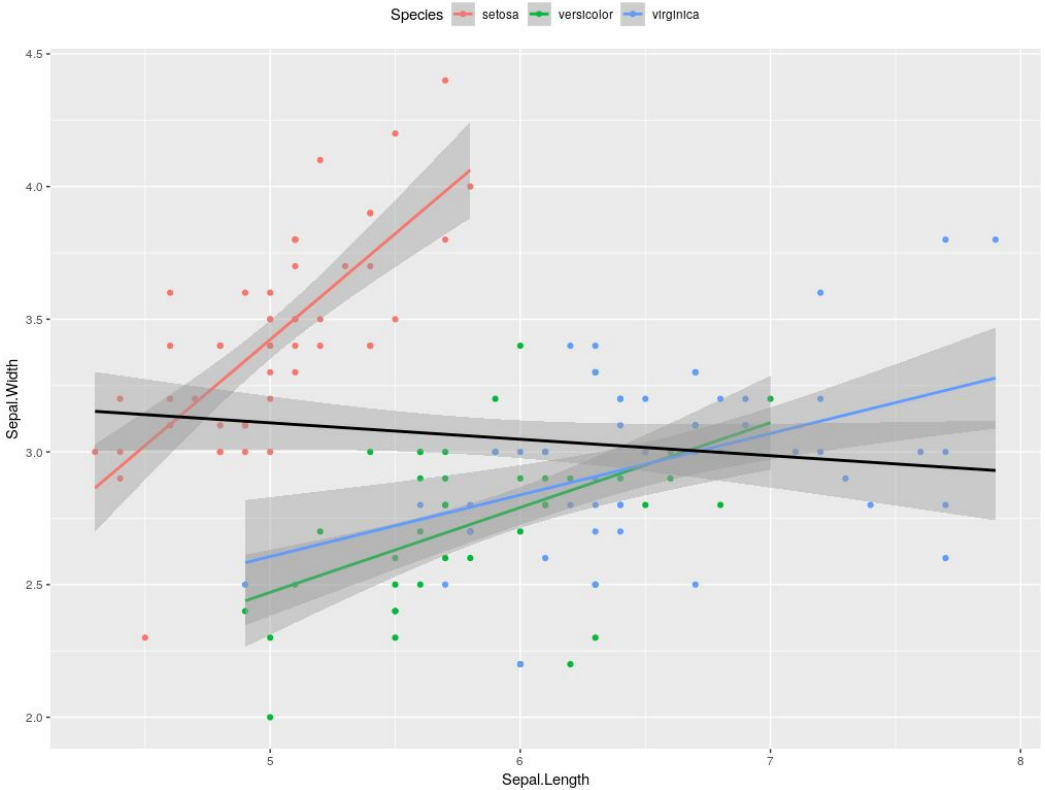
	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**



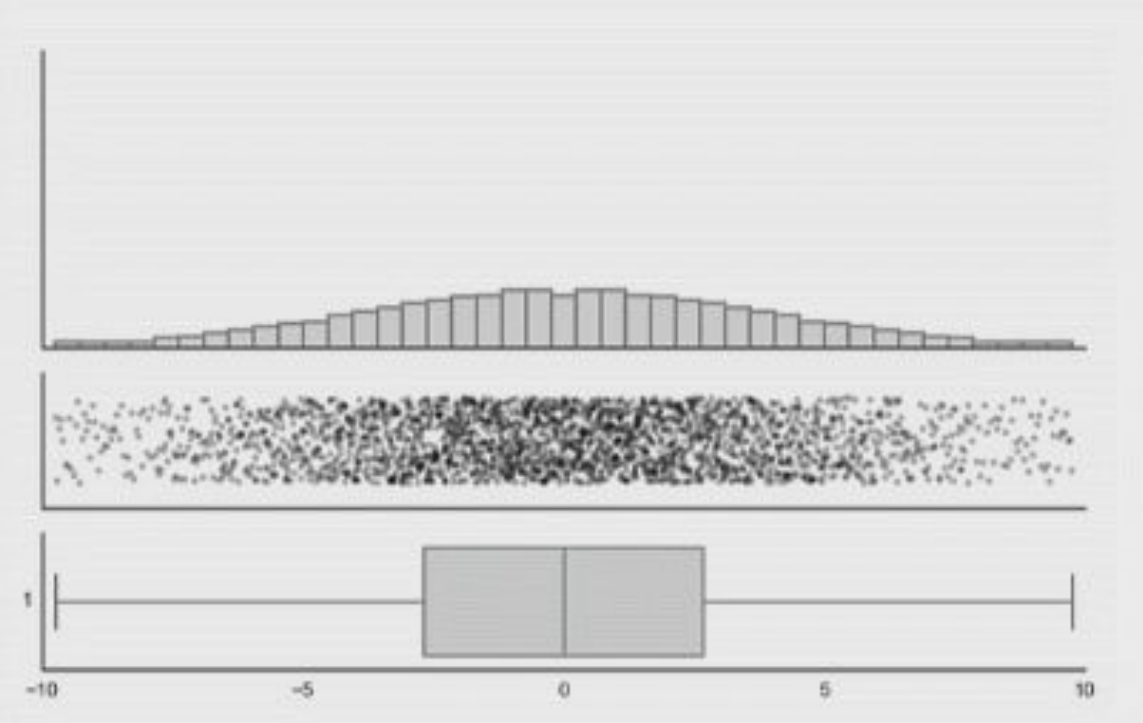
Bickel (1975): Sex Bias in Graduate Admissions: Data From Berkeley

Example Datasets (Simpson's Paradox)

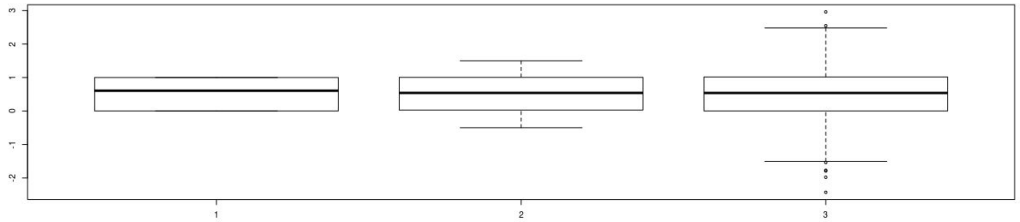


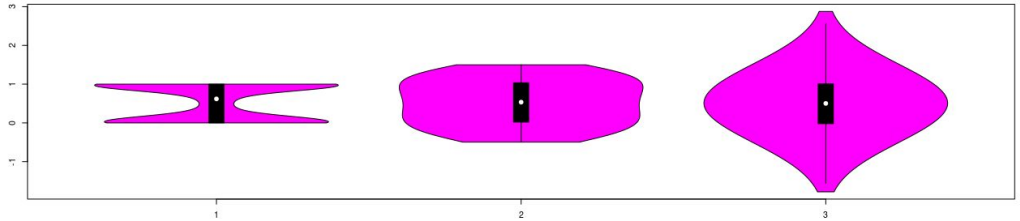

```
ggplot(
  iris,
  aes(
    Sepal.Length,
    Sepal.Width)) +
  geom_point(
    aes(color = Species)) +
  geom_smooth(
    aes(color = Species),
    method = 'lm') +
  geom_smooth(
    method = 'lm',
    color = 'black') +
  theme(legend.position = 'top')
```

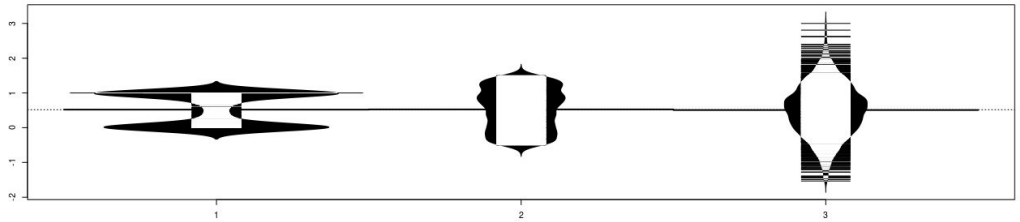
Example Datasets (1D Boxplots)



Example Datasets (1D Boxplots)





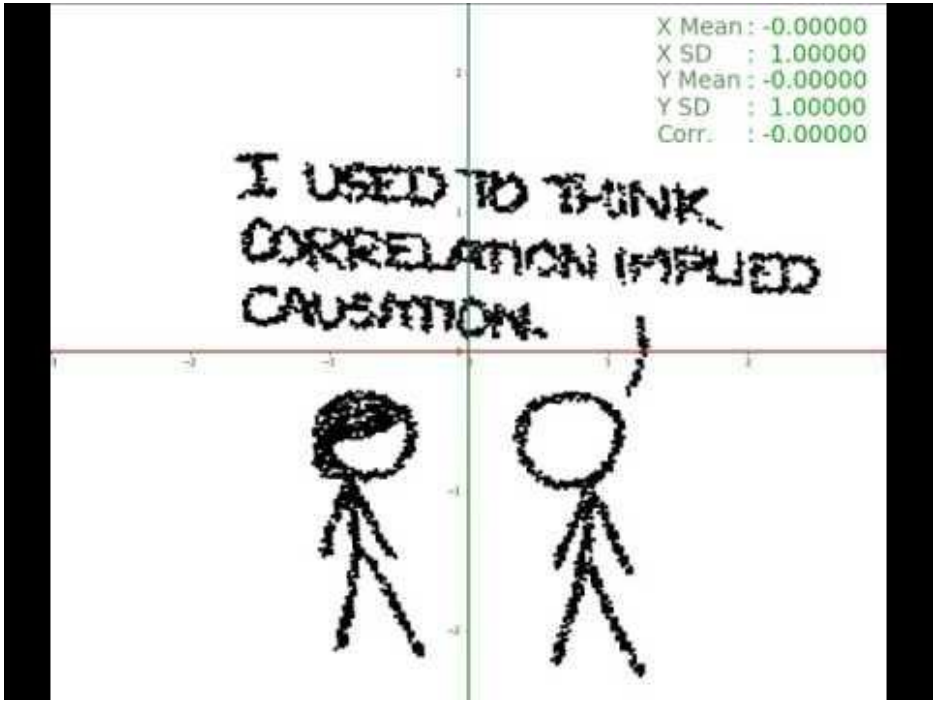


```r
boxplot(
  rbeta(1e3, 0.1, 0.1),
  runif(1e3) * 2 - 0.5,
  rnorm(1e3, 0.5, 0.75))

vioplot::vioplot(
  rbeta(1e3, 0.1, 0.1),
  runif(1e3) * 2 - 0.5,
  rnorm(1e3, 0.5, 0.75))

beanplot::beanplot(
  rbeta(1e3, 0.1, 0.1),
  runif(1e3) * 2 - 0.5,
  rnorm(1e3, 0.5, 0.75))
```

Other Examples



Köszönöm a figyelmet!

Source: https://github.com/tjwei/Animation-with-Identical-Statistics