GENERAL ASSEMBLY

DATA SCIENCE
CLASS 3: DATA VISUALIZATION

AGENDA

I. THE IMPORTANCE OF VISUALIZATION
II. VISUALIZATION AS A MEDIUM

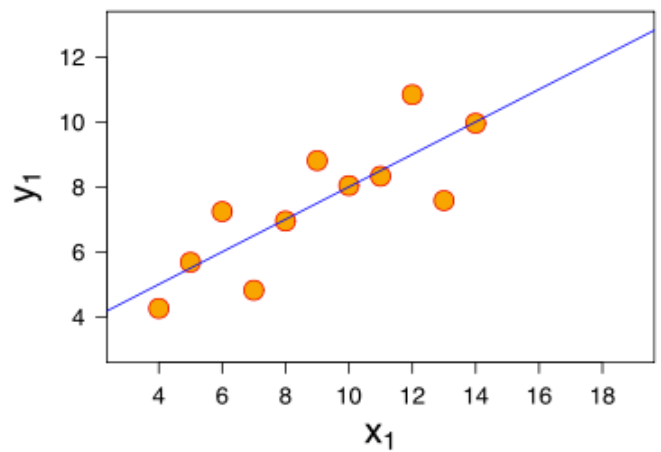
LAB:
III. VISUALIZATION IN R

I. THE IMPORTANCE OF VISUALIZATION

VISUALIZATION VS. SUMMARY STATISTICS

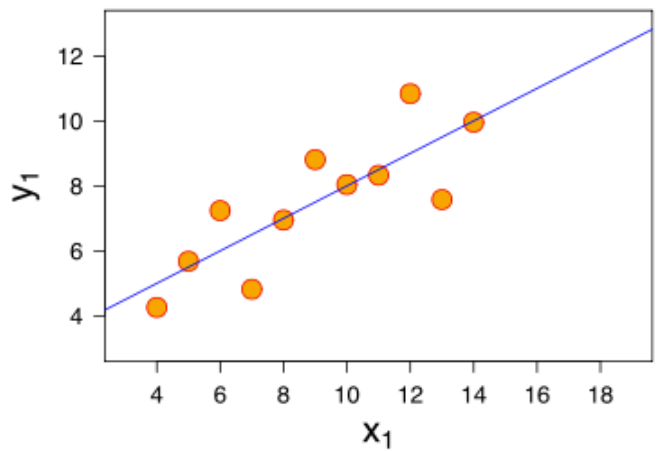
Consider the following dataset:

- eleven (x, y) points



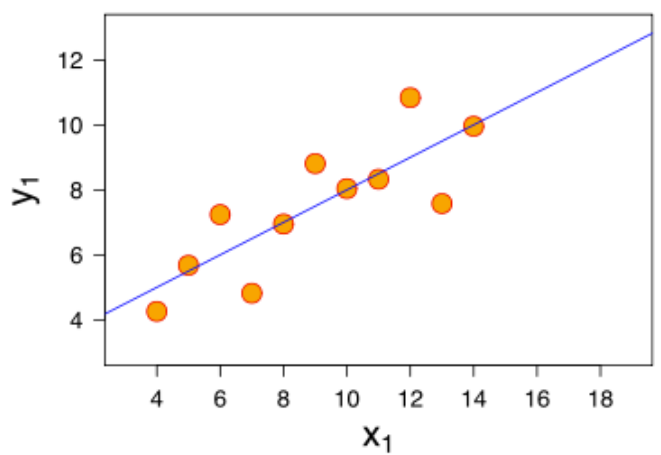
Consider the following dataset:

- eleven (x, y) points
- mean of x = 9, mean of y = 7.5



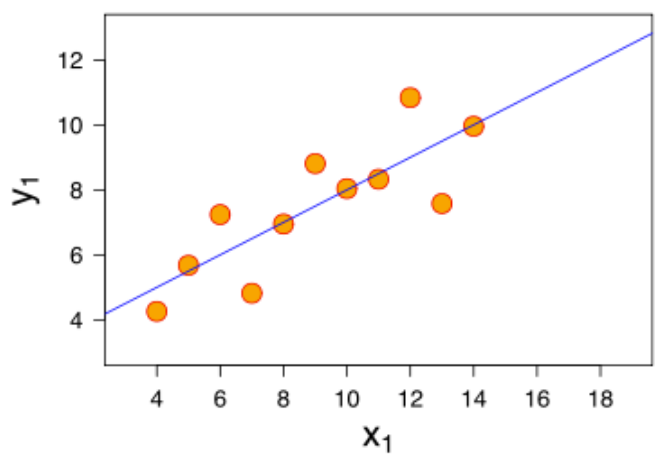
Consider the following dataset:

- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1



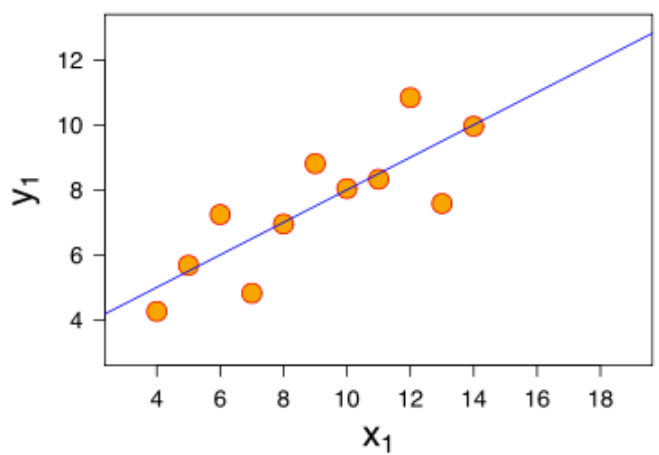
Consider the following dataset:

- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1
- correlation of x and y = 0.8



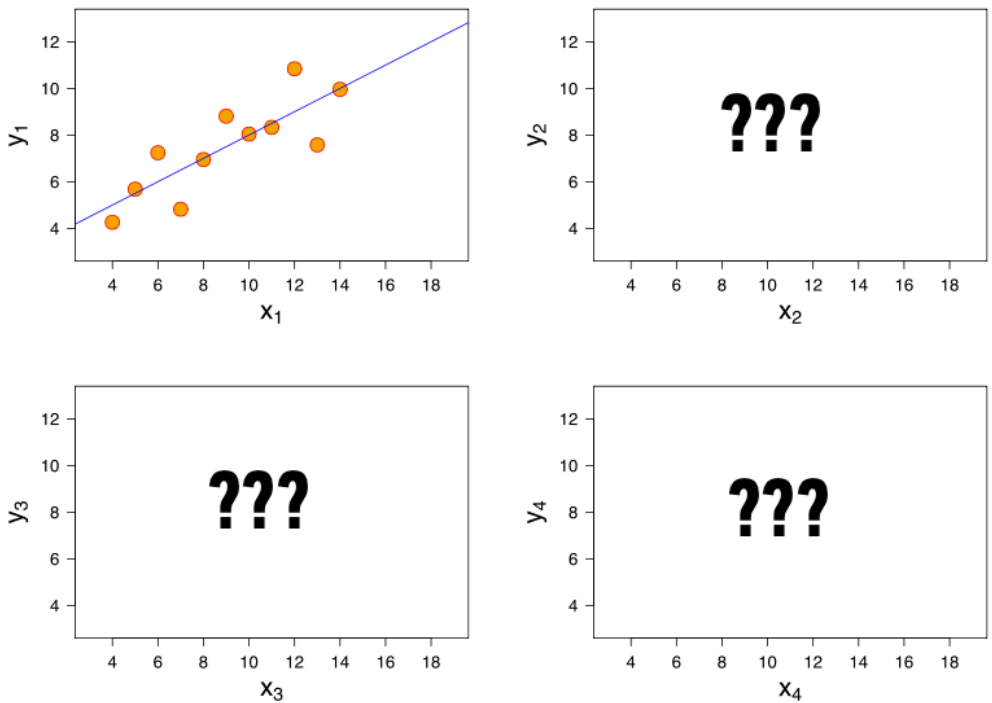
Consider the following dataset:

- eleven (x, y) points
- mean of x = 9, mean of y = 7.5
- variance of x = 11, variance of y = 4.1
- correlation of x, y = 0.8
- line of best fit: y = 3.00 + 0.500x



Now, suppose I give you three more datasets with exactly the same characteristics...

Q: how similar are these datasets?

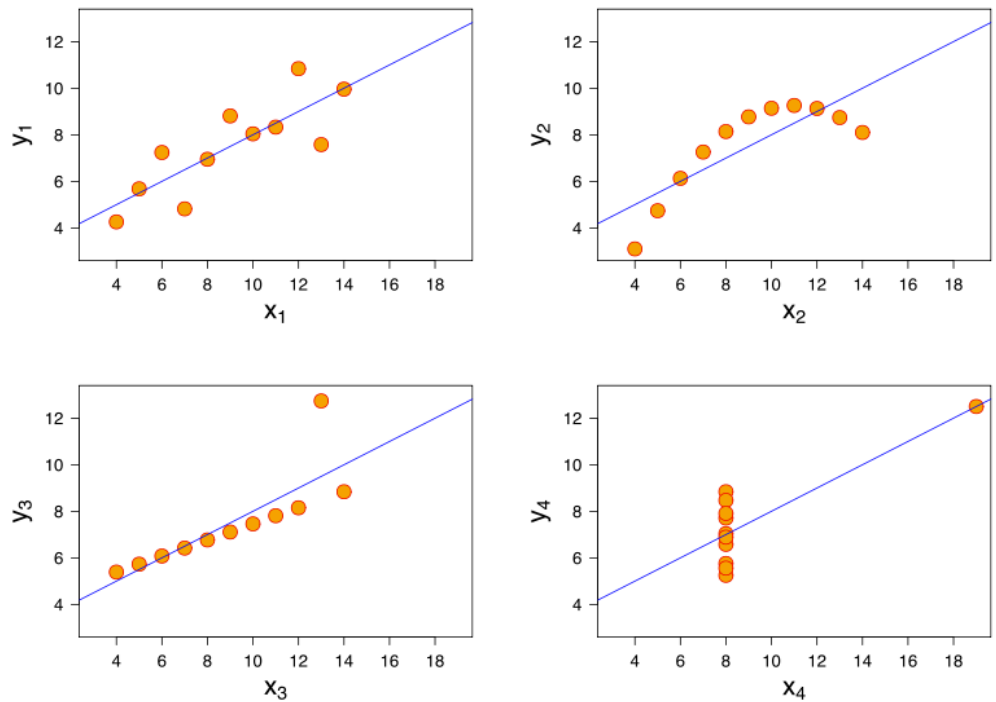


Now, suppose I give you three more datasets with exactly the same characteristics.

Q: how similar are these datasets?

A: not very!

http://en.wikipedia.org/wiki/Anscombe's_quartet

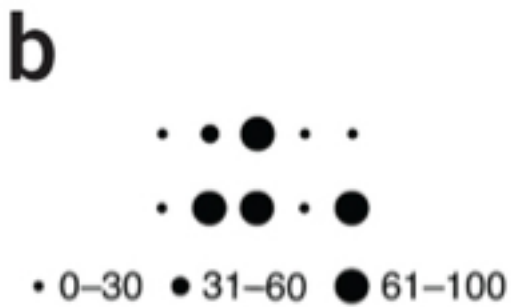


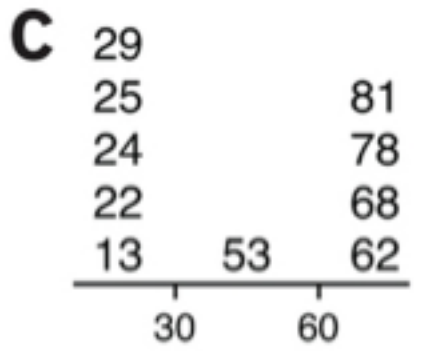
VISUALIZATION INTERACTS WITH INTERPRETATION

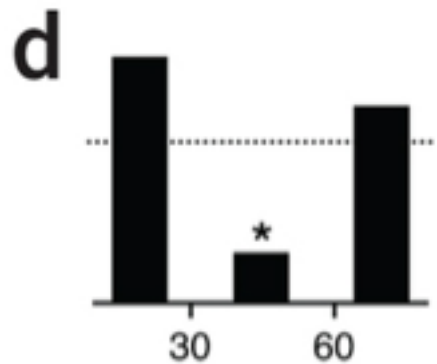
a

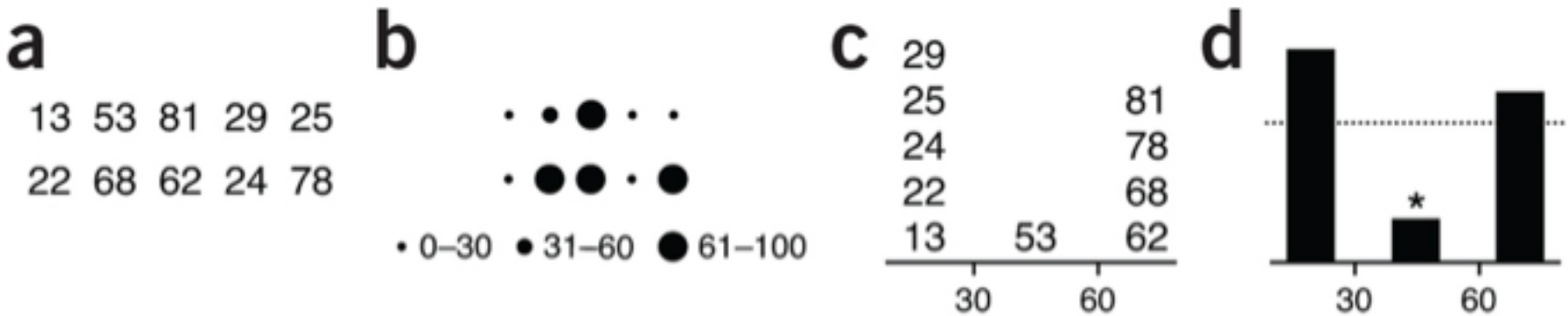
13 53 81 29 25

22 68 62 24 78


b
0–30
31–60
61–100

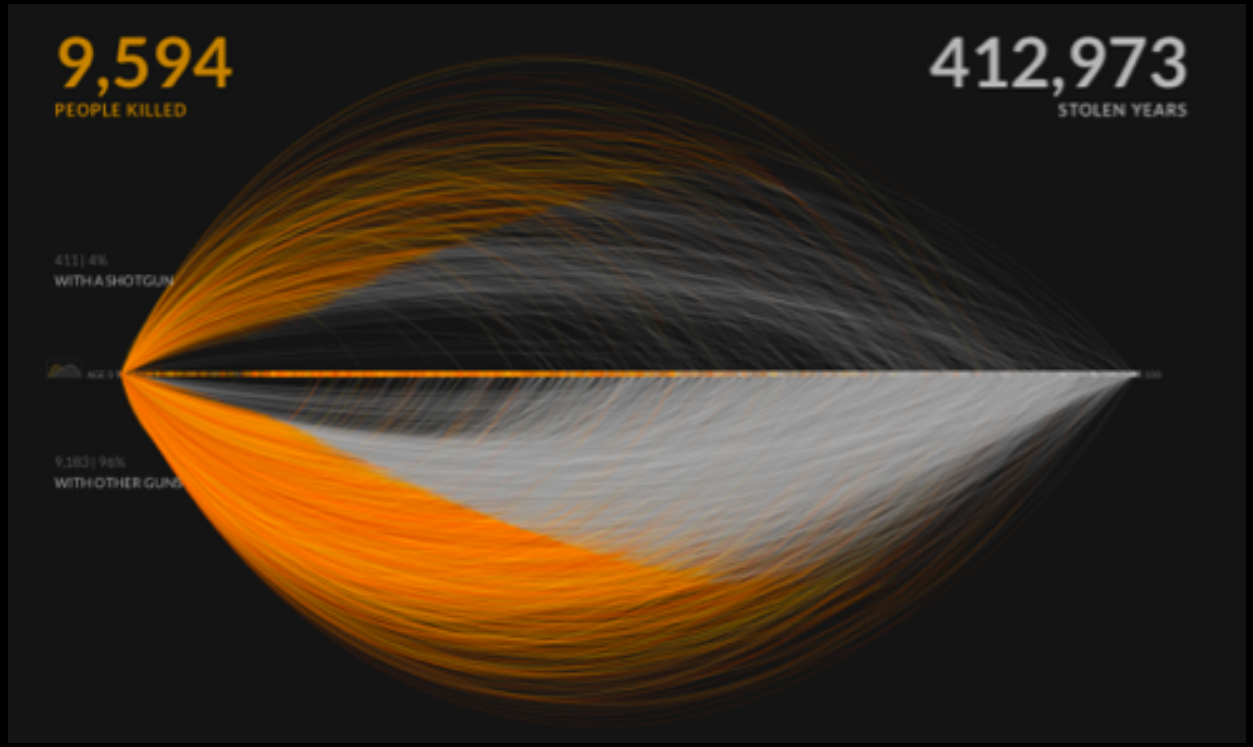

C
29
25
24
22
13
81
78
68
62
53
30
60


d
*
30
60


a
13 53 81 29 25
22 68 62 24 78
b
• 0–30 • 31–60 • 61–100
c
29
25
24
22
13
81
78
68
62
53
30
60
d
*
30
60

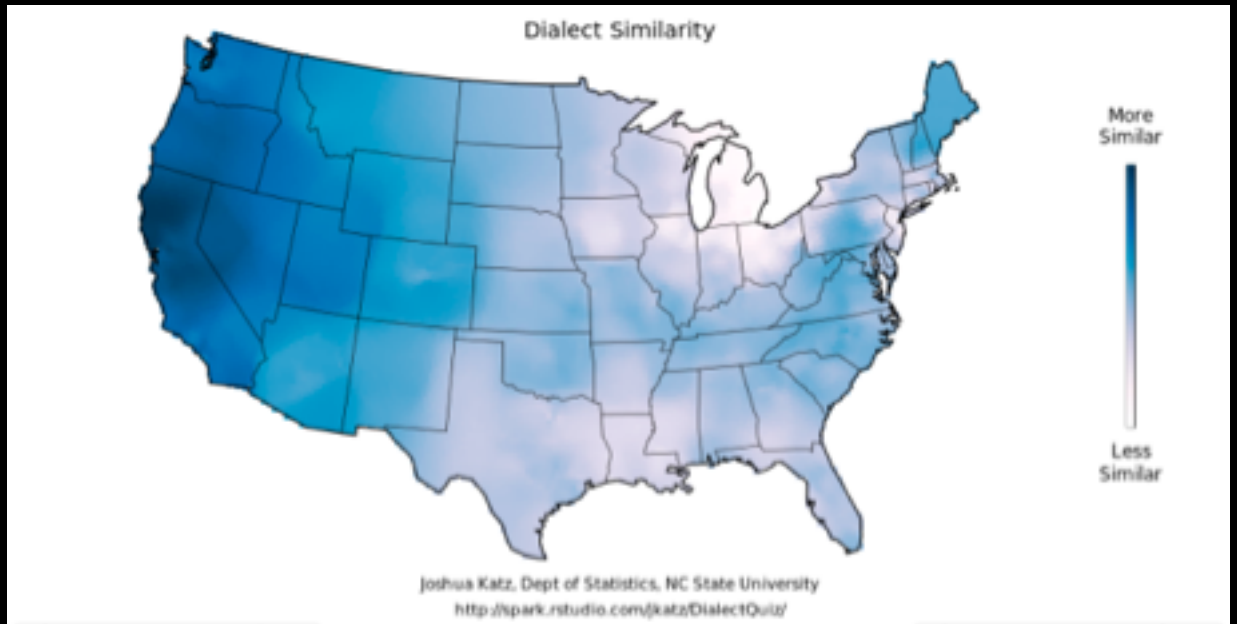
II. VISUALIZATION AS A MEDIUM

VISUALIZATION AS A MEDIUM



VISUALIZATION AS A MEDIUM

VISUALIZATION AS A MEDIUM



	Most Similar Cities	
1	Santa Rosa CA	59.5
2	Vallejo CA	59.2
3	Fremont CA	58.9
4	San Francisco CA	58.8
5	Modesto CA	58.0

	Least Similar Cities	
1	Detroit MI	33.7
2	Warren MI	34.1
3	Toledo OH	35.0
4	Flint MI	35.1
5	Lansing MI	35.1

III. VISUALIZATION IN R