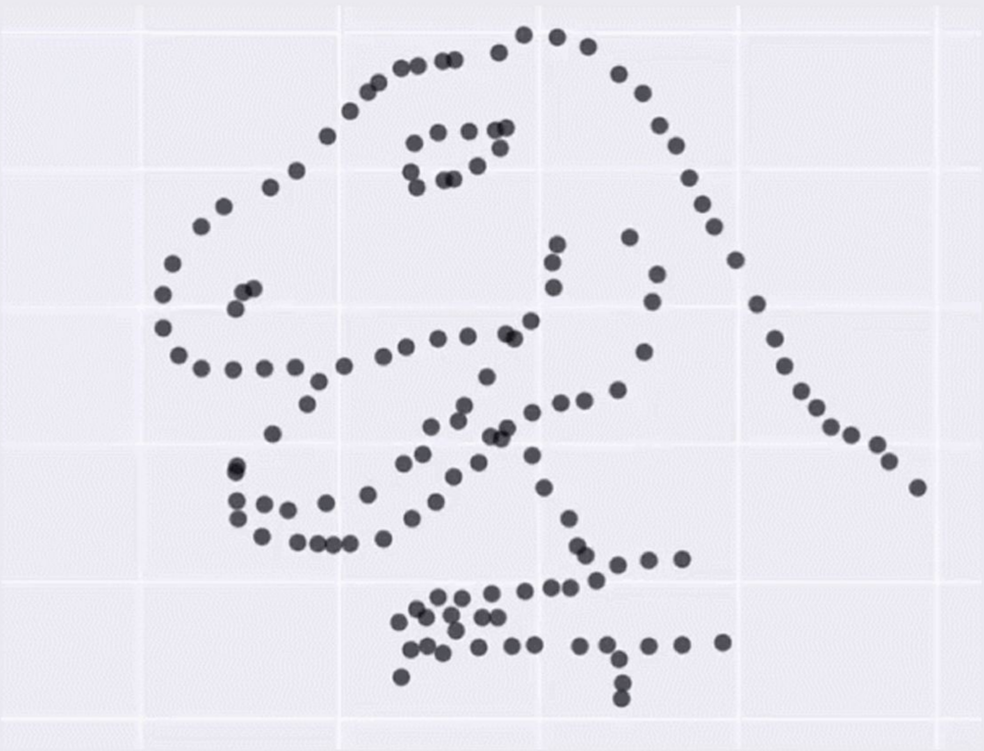
Statistics won't tell you about the T-Rex



Why is visualization so important during data analysis?

Graphs are superfluous if you have **statistics** to evaluate your data right?

Short answer: **wrong.**

The limits of statistics were very effectively demonstrated in 1973 by the statistician Francis Anscombe.

Let’s have a look at his demonstration.

Look at the following data dable.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Let’s analyze the data statistically.

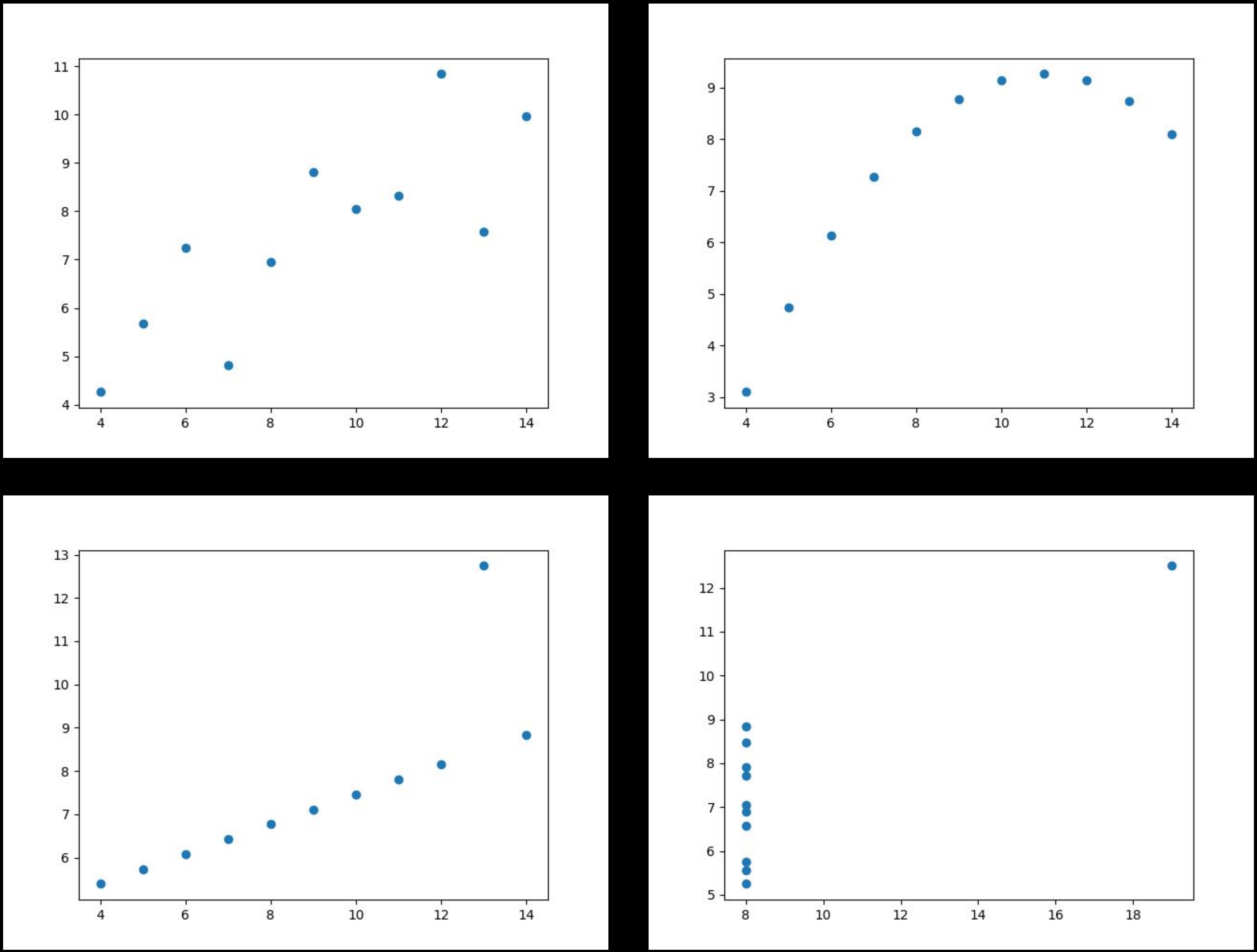
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Each dataset has almost the same statistics.
The mean and standard deviation for X are 9 and 3.32.
For Y the mean and standard deviation are 7.5 and 2.03.

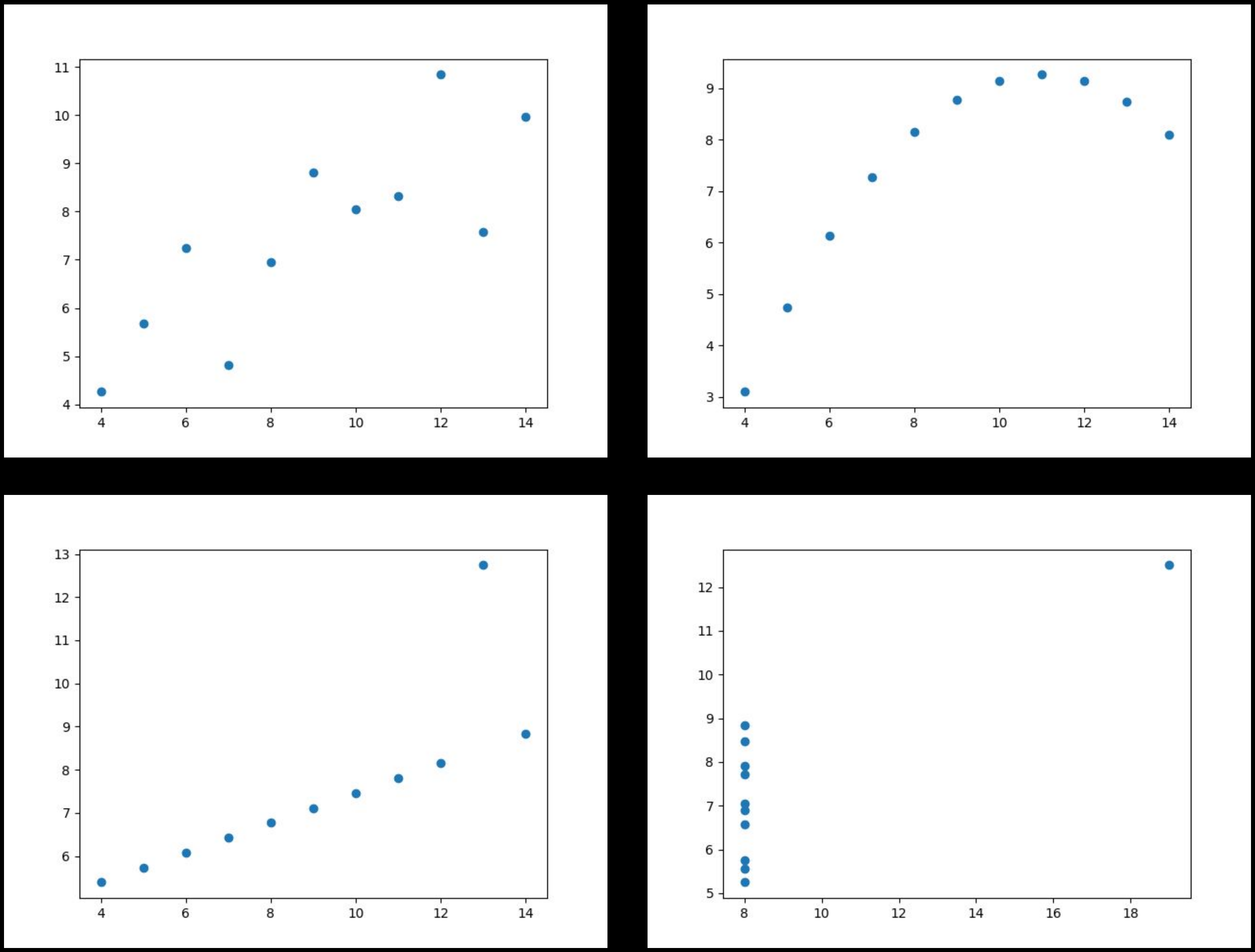
With a value of 0.82 the pearson correlation between X and Y is identical to two decimals places for all four data sets as well.

So we might just say the datasets are identical.

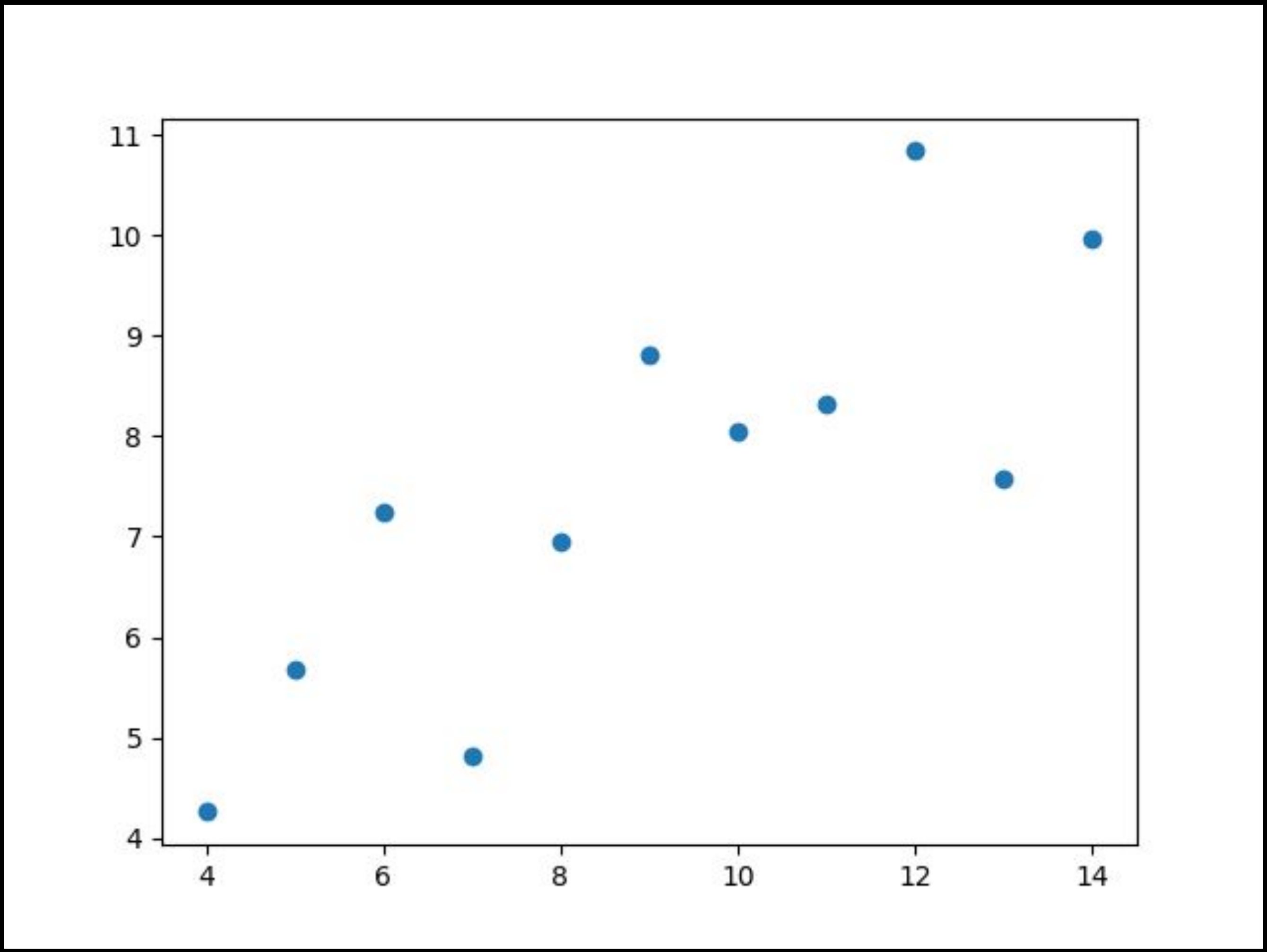
Let’s see if we can confirm that there is no real difference after plotting the data.



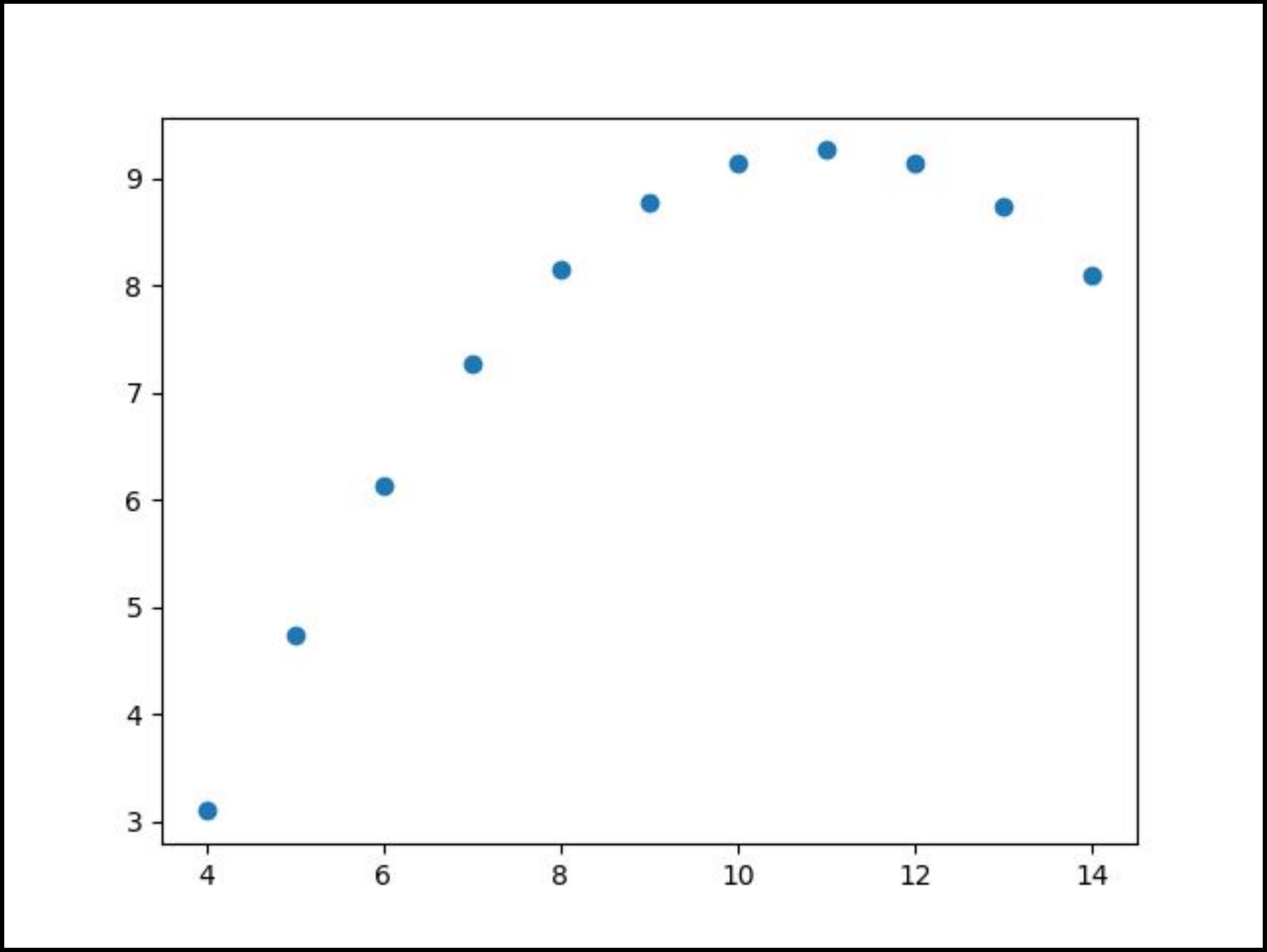
Suddenly we can see many differences between the different data sets.



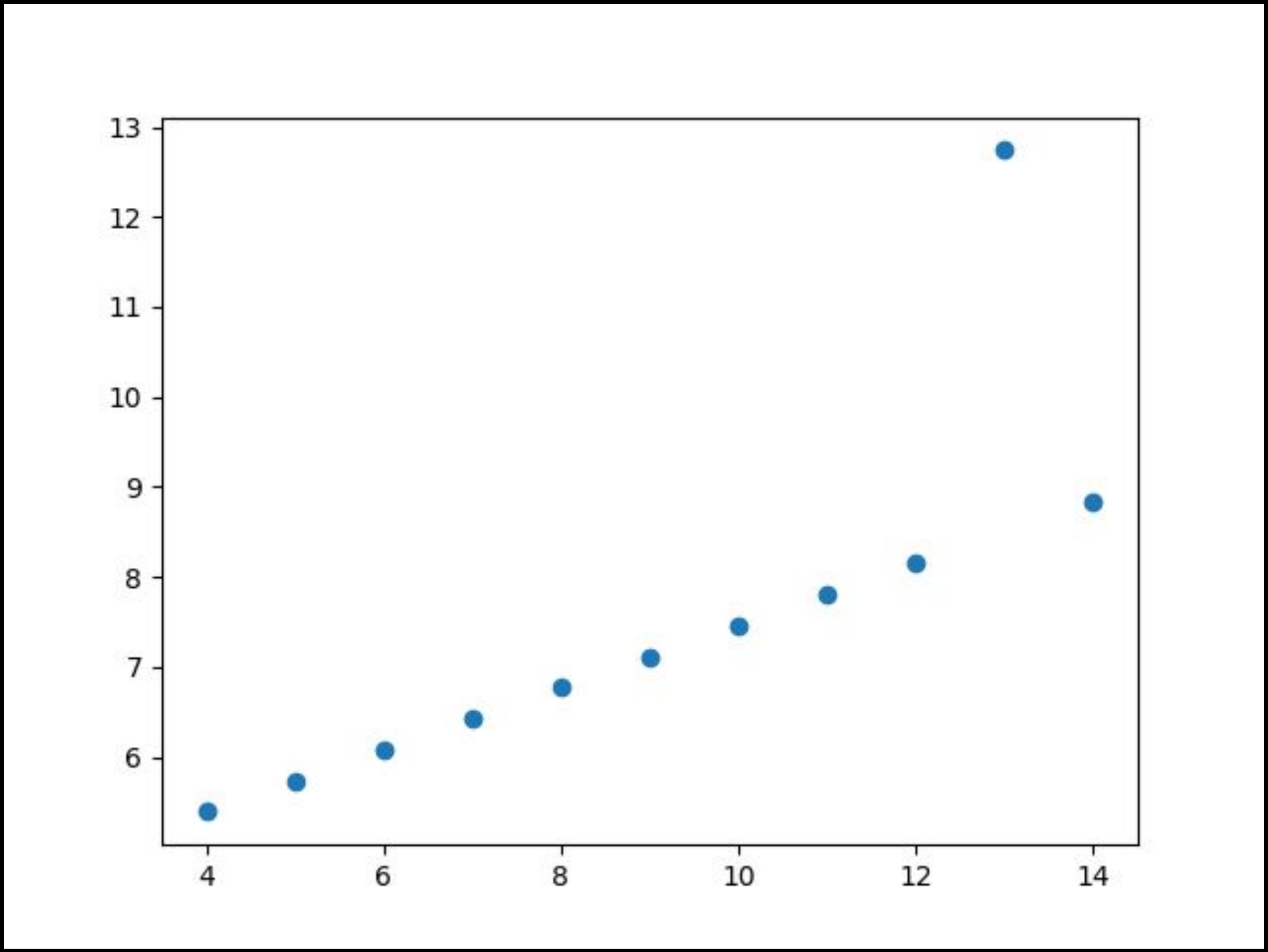
The first data set seems to show a linear relationship between X and Y.



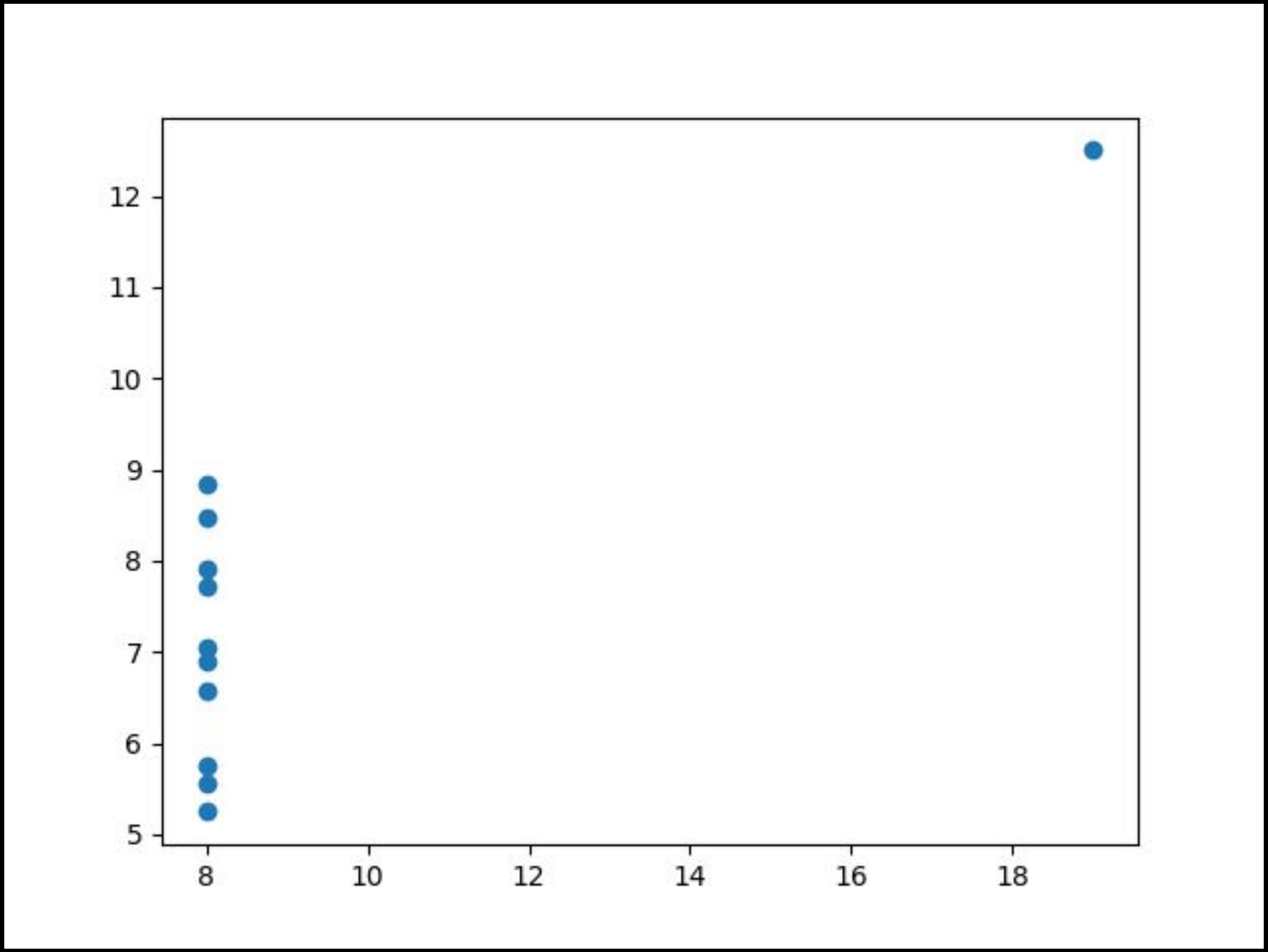
In the second dataset there is a non-linear relationship between X and Y.



The third data set shows a perfect linear relationship between X and Y, with one outlier.



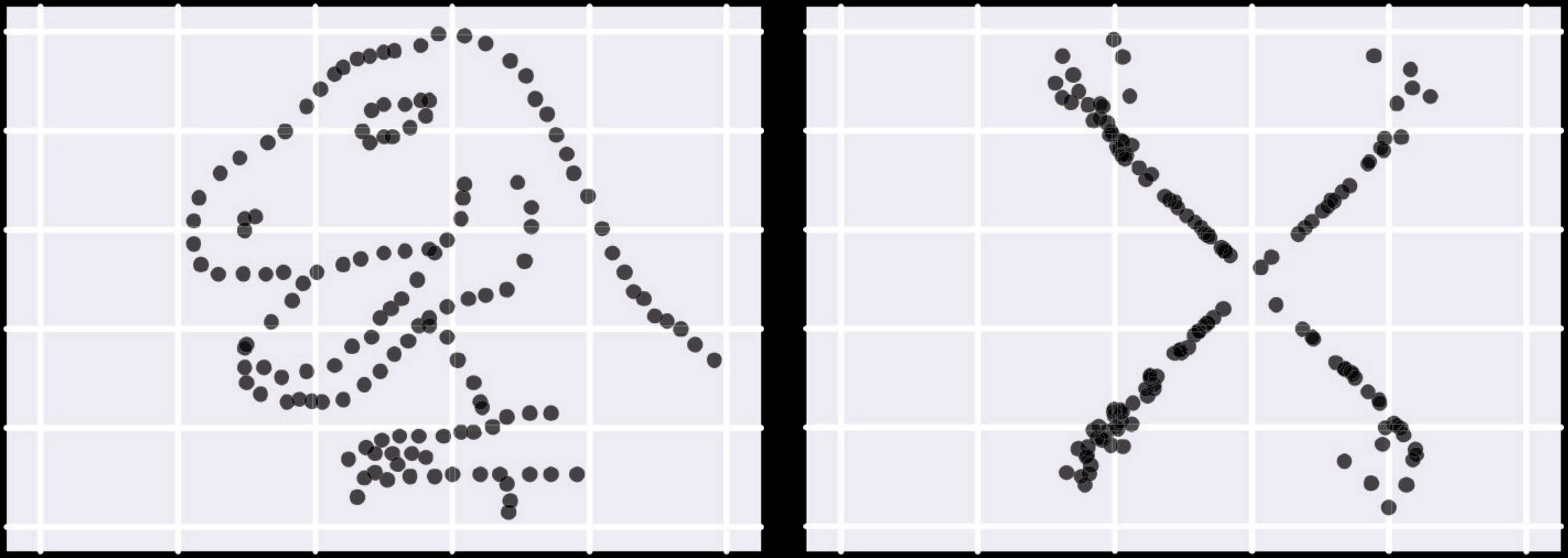
In the fourth dataset one high-leverage point produces a high correlation coefficient.



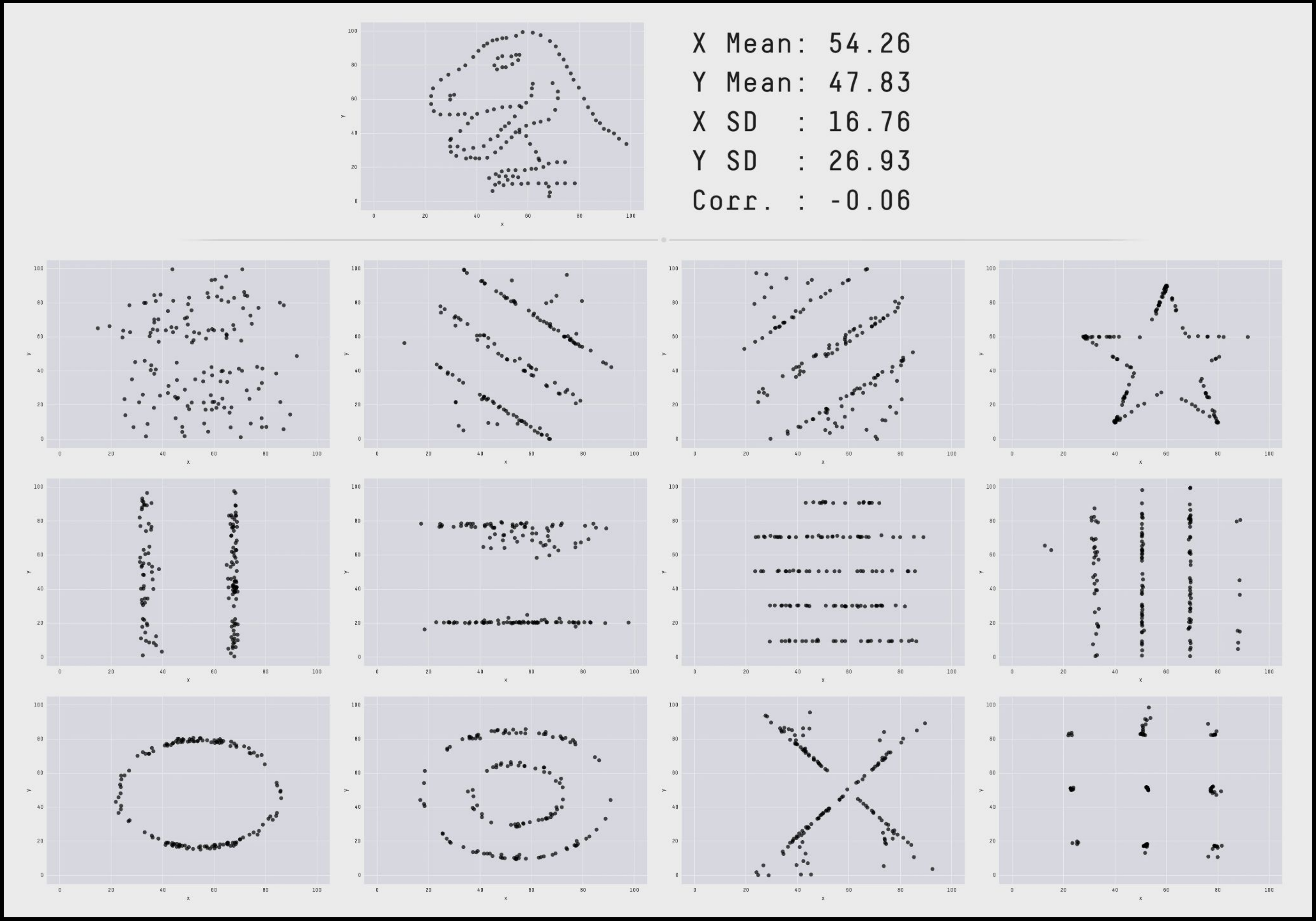
It doesn’t end here.

In a 2017 research paper, Justin Matejka and George Fitzmaurice presented a technique that can generate data with different appearance but identical statistics.

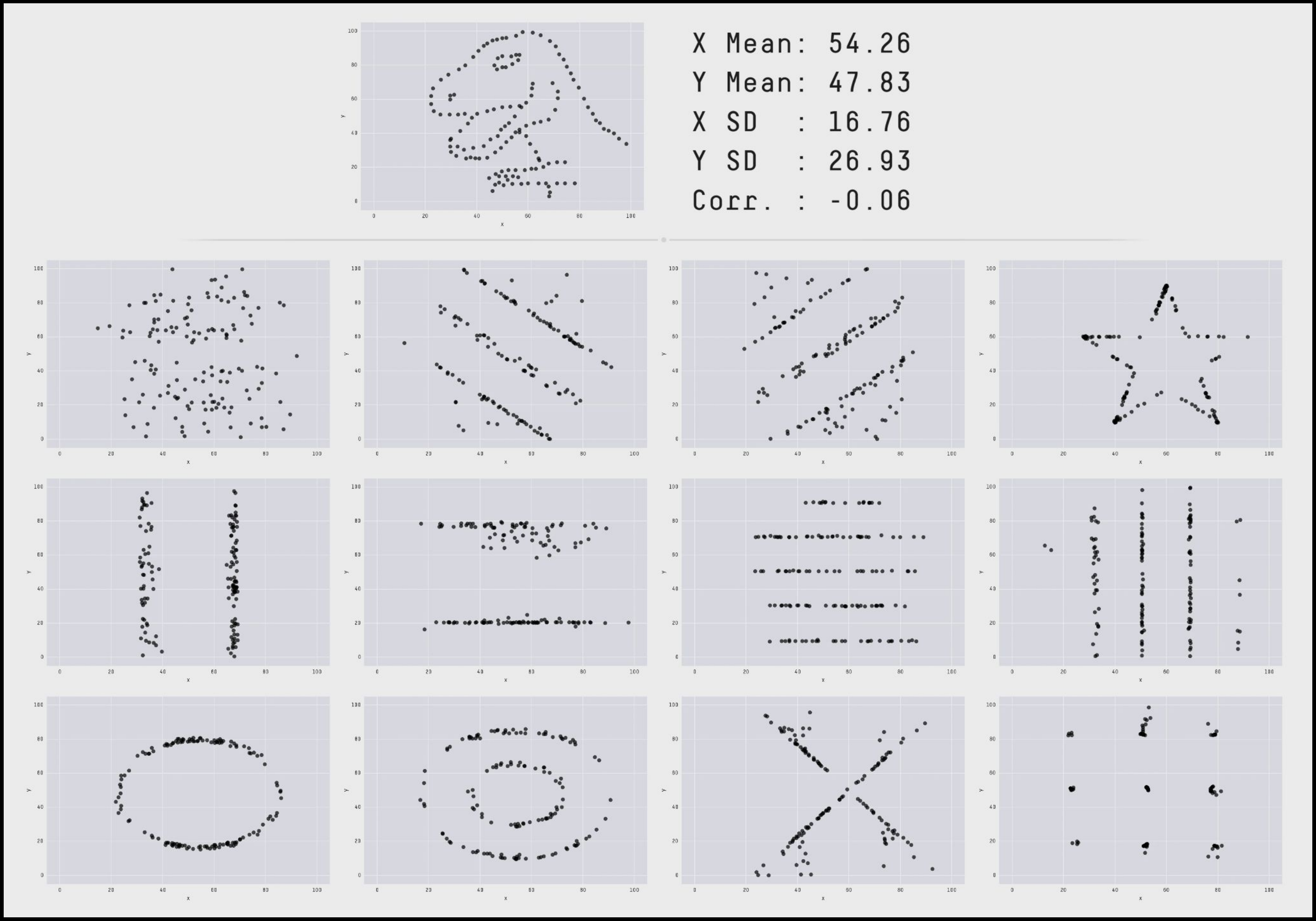
Would you have thought that these two datasets share the same statistics? But it gets even better!



These 13 datasets generated with this approach have identical statistics.



But they look completely different and are based on input sketches!



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

One consequence of this research is that **data can be easily manipulated** without being detected by statistical methods.

This highlights the importance of proper data cleaning and analysis.

Data engineers and analysts, rest assured: Your work is very, very important.

You are there to show us the T-Rex, when statistics fails.

Remember

1. Statistics might not reveal the T-Rex hiding in your data.

2. Data manipulation might not be detectable with statistical methods alone.

3. Data analysis should utilize visualization in addition to statistical analysis.

Feel free to reach out or to connect with me for more weekly slideshows on visualization, data science and machine learning.