# Why Data Visualizations Are Not Optional: Anscombe's Quartet

This course is not intended to have any theory. It's a practical, hands-on exploration of visualization and storytelling. But there's some things that can't be avoided. One of these concepts is what's known as the Anscombe's Quartet.

Francis Anscombe was a prolific statistician at Yale University. He developed fundamental concepts around residuals in linear regression and subjective probability which advanced the study of economics.

But to us doing data visualization, he helped to overturn the premise that "numerical calculations are exact, but graphs are rough." In reality both are important but visualizations can "help us perceive and appreciate some broad features of the data" and allow us to "look behind those broad features and see what else is there."[1] He said this in 1973, well before the revolution in computing power that we see today.

Anscombe constructed four datasets (I – IV) that have an x and a y:

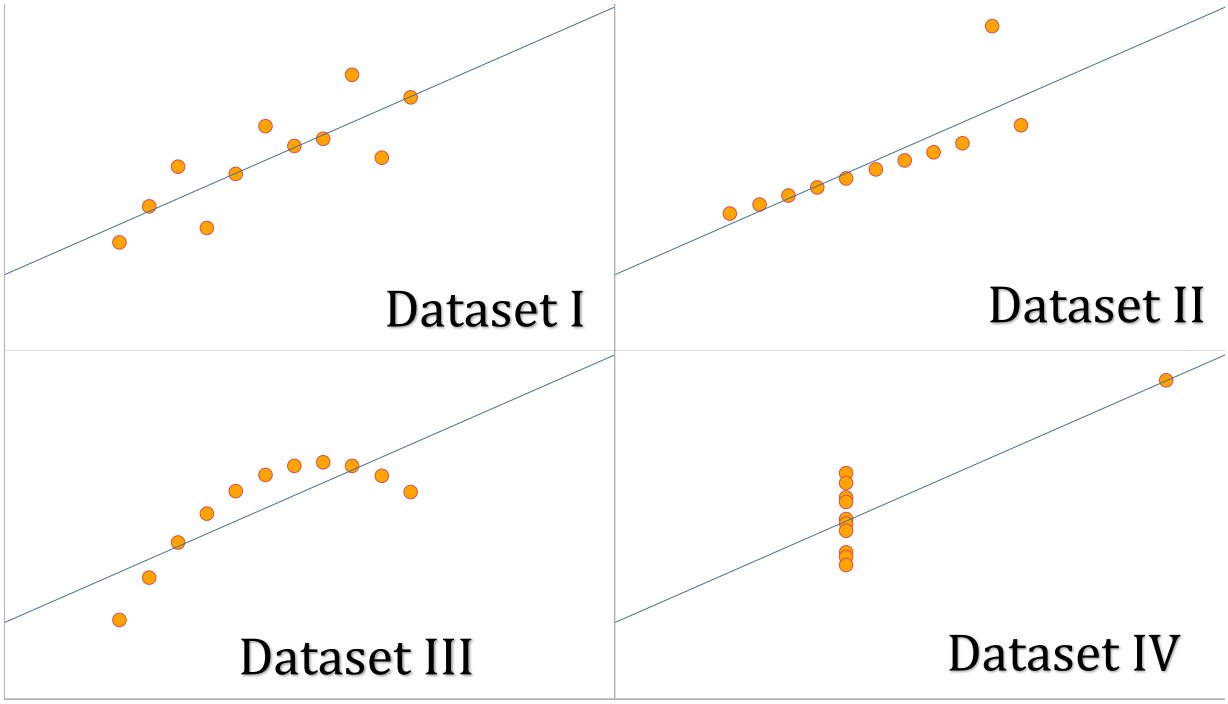
| **I** | | **II** | | **III** | | **IV** | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

[1] Anscombe, F. J. "Graphs in Statistical Analysis." The American Statistician 27, no. 1 (1973): 17-21.

Each dataset would ALL have the following properties:

1. The mean of x is 9, the mean of y is 7.5;
2. The variance of x is 11, the variance of y is 4.12.
3. The correlation between x and y is identical and is 0.816.
4. The equation for a linear regression for each dataset is the same— $y = 0.5x + 3$.
5. The $R^2$ of the regressions are all 0.667.
6. The standard error of all of the datasets is 0.118.

If you present the numbers in a nice table, you will show a solid set of summary statistics. However, when you plot x, y and the fitted line, $y = 0.5x + 3$, you will see something dramatically different as see in the visuals below.




So what do these visuals show?

1. Dataset I shows the most common type of scatterplot with no outliers.
2. Dataset II shows one outlier that nudges the regression line slope up to 0.5.
3. Dataset III is not linear even though a linear regression line could be fitted.
4. Dataset IV has an outlier that can produce a high correlation coefficient in an otherwise nonlinear set of data.[2]

## Can This Be Shown In Tableau?

Yes. Just replicate how I did it in this Tableau file (https://ucdavis.box.com/s/nhdrw08uftgqyr3f8njiuxuzwtw6z380).

[2] As an aside, dataset IV can be used as a visual proof of the common adage "correlation is not causation".