Bits & Bytes: Five Tips on Correlation

|  |
| --- |
| REF: <https://twitter.com/altonncf/status/293392615225823232> |
| A few years ago, one of my classmates showed me this chart from Twitter, and we laughed a lot about this witty Chrome PR. Though I do not exactly remember, I assume it was just an Internet joke rather than actual presentation by Chrome.  CORRELATION is one of the words that is commonly used in our business and personal life, and most people understand the meaning of the term 95% correctly. But, I would like to talk about the 5% part here – for fun ^^.  What do you think about the misconception of this chart?  Yes, you might heard about the phrase:  Correlation does not imply causation!  Correlation is used for showing the strength of a relationship between two variables. In this example, Internet Explorer (IE) market share and the Murder Rate have the same trend as both of the two phenomena decrease over time – more interestingly, with the same pattern.  However, we cannot tell the decline of Murder Rate is caused by the compression of the IE market share (or vise versa). They might have some very different reasons for this similar trend. |
| https://upload.wikimedia.org/wikipedia/commons/thumb/e/ec/Anscombe%27s_quartet_3.svg/325px-Anscombe%27s_quartet_3.svg.png  REF: <https://en.wikipedia.org/wiki/Correlation_and_dependence> |
| Next, I would like to introduce these four plots above and ask you if you can find any common characteristics among them. You might be surprised that they share five common factors:   * They all have 11 data points. * Y variables have the same mean of 7.5. * Their scale of diversity within y are the same - variance of 4.12. * When we fit the line for these four datasets, they have the same line - say,   Linear regression line with y = 3 + 0.5x.   * They have the same correlation coefficient of 0.816.   Continuing to talk about correlation, only the first plot has the valid correlation coefficient. This is because Correlation represents Linear Relationships only.  It is absolutely fine to use the term CORRELATION to express any association between two subjects in general terms however correlation as a technical term does not explain the curve like the second plot. The third plot’s correlation could be 1 (means perfectly correlated) if the data point on the top has been removed. The last plot could have a correlation of 0 (meaning no-correlation at all) if the data point on the top-right corner has been removed – those points are called OUTLIERS.  These imply that Correlation can be disfigured by outliers. |
| REF: <https://www.weforum.org/agenda/2016/03/does-working-fewer-hours-make-you-more-productive/> |
| After we have had a long weekend most of us feel fresh and dream to have regular 3-day weekends. It might be ideal to have a 4-3 ratio work life balance so the above graph is a bit of good news. The plot above presents working hours vs. GDP created per person in the OECD between 1990 and 2012. It tells us that working shorter hours results in more productive labour outcomes.  So, we can say that Negative correlation is not always bad. |
| Lets look at the plot on the left – do you think there is any relationship between dataX and dataY?  Yes, these two data sit perfectly on a circle - it means that  , and we cannot deny a strong relationship between dataX and dataY.  So how much correlation do you think they have?  Ok, it is 0(zero)! – Zero because, correlation (technically) only counts the LINEAR relationship, as mentioned earlier.  But,  Correlation of zero can hide some secret relationship behind. |
| Hope you all value this extra 5% of knowledge about Correlation and make full use of it for the business. |